

Hadoop Administrator Interview Questions

Cloudera® Enterprise version Include 155 Interview Questions

About book

Cloudera Enterprise is one of the fastest growing platforms for the BigData computing world, which accommodate various open source tools like CDH, Hive, Impala, HBase and many more as well as licensed products like Cloudera Manager and Cloudera Navigator. There are various organization who had already deployed the Cloudera Enterprise solution in the production env, and running millions of queries and data processing on daily basis. Cloudera Enterprise is such a vast and managed platform, that as individual, cannot manage the entire cluster. Even single administrator cannot have entire cluster knowledge, that's the reason there is a huge demand for the Cloudera Administrator in the market specially in the North America, Canada, France, UAE, Germany, India etc. Many international investment and retail bank already installed the Cloudera Enterprise in the production environment, Healthcare and retail e-commerce industry which has huge volume of data generated on daily basis do not have a choice and they have to have Hadoop based platform deployed. Cloudera Enterprise is the pioneer and not any other company is close to the Cloudera for the Hadoop Solution, and demand for [Cloudera certified Hadoop Administrators](#) are high in demand. That's the reason HadoopExam is launching Hadoop Administrator Interview Preparation Material, which is specially designed for the Cloudera Enterprise product, you have to go through all the questions mentioned in this book before your real interview. This book certainly helpful for your real interview, however does not guarantee that you will clear that interview or not. In this book we have covered various terminology, concepts, architectural perspective, Impala, Hive, Cloudera Manager, Cloudera Navigator and Some part of Cloudera Altus. We will be continuously upgrading this book. So, you can get the access to most recent material. Please keep in mind this book is written mainly for the Cloudera Enterprise Hadoop Administrator, and it may be helpful if you are working on any other Hadoop Solution provider as well.

Feedback

This is a full-length book from <http://hadoopexam.com> and we love the feedback so that we can improve the quality of the book. Please send your feedback on hadoopexam@gmail.com or admin@hadoopexam.com

Even you want to share your experience and story with the preparation of the real exam certification, interview please share the same. It would help other candidates as well.

Restrictions

Entire content of this book is owned by HadoopExam.com and before using it or publishing anywhere else either digitally on web or printing and distribution require prior written permission from HadoopExam.com. You **cannot** use the code or exercises from this book in your software development or in your software product (commercial as well as open source) and you need to take prior written permission to use the same.

Copyright© Material

This book contents are copyright material and it is hard work and many years of experience working with disruptive technologies, which helps in producing this material. All rights are reserved on the material published in this book. You are not allowed to any part of this material to be reproduced, stored in a retrieval system, and must not be transmitted in any form or by any means, without the prior written permission of the author and publisher, except in the case of brief quotations embedded in critical articles or online and off-line reviews. Wherever, you use contents make sure full detail of the book is mentioned.

Author had tried as much as his/her capacity in preparing of this book so that accuracy can be maintained in the presented material. The material sold using this book does not have any warranty or guaranty either express or implied. Neither of the author, publisher, dealer and distributors will be held liable and responsible (explicit/implicit these all parties mentioned are not liable and responsible) for any damages caused or alleged to be caused directly or indirectly by this book. You should note, this material is part of your learning process and as time passes material can be outdated and you should wait or look for that latest material.

Author and publisher have endeavored to provide trademark information about all of the companies and products mentioned in this book. However, we cannot guarantee the accuracy of this information.

Disclaimer:

1. Hortonworks® is a registered trademark of Hortonworks.
2. Cloudera® is a registered trademark of Cloudera Inc
3. Azure® is a registered trademark of Microsoft Inc.
4. Oracle®, Java® are registered trademark of Oracle Inc
5. SAS® is a registered trademark of SAS Inc
6. IBM® is a registered trademark of IBM Inc
7. DataStax® is a registered trademark of DataStax
8. MapR® is a registered trademark of MapR Inc.
9. Apache® is a registered trademark of Apache Foundation
10. Databricks® is a registered trademark of Databricks Inc

Publication Information

First Version Published: Jan 2020

Edition: Early

Piracy

We highly discourage the piracy of copyright material especially it happened online on the internet. Piracy causes the damages to all, first of all it damages yourself by not honestly using the correct material, generally pirated material is edited and wrong information is presented which can make big damage as part of your learning process. As well as when you become author and honestly write similar

material, piracy will damage your material with the same extent or more. Hence, don't encourage piracy. If piracy is reduced cost of material will automatically decrease. It also makes damages to author, publisher, dealer and distributors. If you come across any illegal copies of this work in any form on the Internet, then please share the detail with URL, location or website name immediately on email id hadoopexam@gmail.com we really appreciate your help in protecting author's hard work and also help in reducing the cost of material.

Author/Trainer required

Corporate Trainer: We have many requirements, where our corporate partners need their team to be trained on particular skill sets. If you are already providing corporate trainings for any skill set, then please become our onsite training partner and fill in the form mentioned above and our respective team will contact you soon. You will get very good revenue for sure. However, what we want, you must be able to train our corporate partner resources. What matters to us? Your proficiency in a particular domain/skill and good oral communication skills. You must be able to be accessible to learners as well.

Online Trainer: If you are a working professional and master or proficient in any particular skills and feel that, you are capable of giving online virtual trainings e.g. 2 hrs a day until course contents are completed. Please send an email at admin@hadoopexam.com. You will get a very good revenue share for sure. What matters to us? Your proficiency in a particular domain/skill and good oral communication skills. It will certainly not impact your daily work.

Self-Paced Trainings: Ok, you want to work as per your comfortable time and at the same time sharpen your skills. You can consider this option. You can create self-paced trainings on particular domain/skills. Please send an email at admin@hadoopexam.com with us as soon as possible. Before somebody else connects with us for the same skill set. Your commitment is very important for us. We respect your work and we will not sell your work in just \$10 or less to acquire more resources. As we know, it takes a good amount of time and you will provide quality material, so we charge reasonable on that so, you will feel motivated with your work and effort. We respect you and your skill.

Certification Material: You may be already certified professional or preparing for particular certification in a specific domain/skill. So why not use this to make money as well as sharing your effort with other learners globally. Please connect with us by filling form or send email at admin@hadoopexam.com and our respective team will contact you soon.

Author: Yes, we are also looking for authors. Who can write books on a particular technology and what you can get certainly a very good revenue sharing and you can bring the same on your resume or linked in profile to show your excellence? Yes, we are not in need of very good oral communication skills, but good writing skill. However, team will also help you to get work done. Author can be more than one for a particular book. However, we wanted you to be in long relationship. So that you don't just write a single e book, but can create an entire series for a particular domain or skill. Good royalty for sure...

Trending Skills (Not limited these):

Hadoop Spark AWS Cloud Azure Cloud Google Cloud	EMC NetApp VMWare CISCO HP	Adobe Alfresco Apple AppSense AutoDesk	Data Analysis Django Docker Drupal Graphics	Infrastructre Automation Internet of Things (IOT) ISO Development Java Java Script
JQuery Kali Linux Laravel Linux Machine Learning	Mobile Application Development NodeJS Android Angular JS Arduino	IBM Watson IBM BPM WebMethod Gemfire Liferay	Scala Python Java SQL/PLSQL Ruby	SAP SAS Salesforce Oracle Cloud Redhat

Cloudera Certification & Training Material Available on HadoopExam.com are below. Please keep visiting website for that latest products



Cloudera Enterprise

Question-1: What is the Cloudera Enterprise?

Answer: Cloudera Enterprise is a combined solution for the Machine Learning, Analytics, Data Engineering etc. Which include the following solutions.

- Data Warehouse
- Data Science
- Data Engineering
- Operational Database
- Run in Cloud, Multi-Cloud or Hybrid Cloud solutions.

Basically, it's a combination of following 3 things

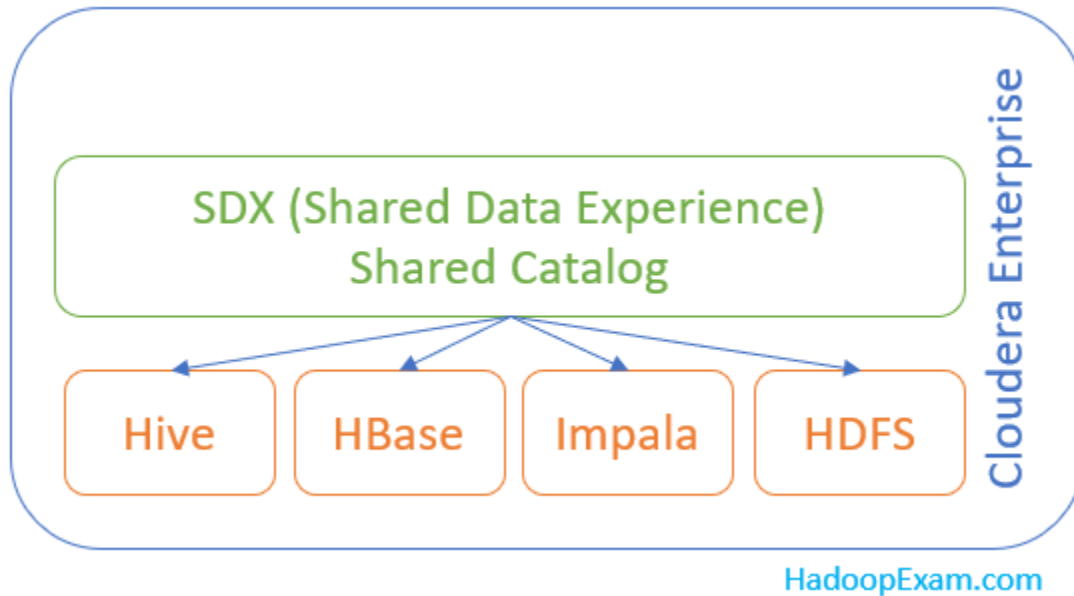
- Open Source CDH (Include Hadoop & its Eco-system)
- Cloudera Manager (Licensed product from Cloudera)
- Cloudera Navigator (Licensed Product from Cloudera)

Question-2: How does Cloudera Enterprise Differ with the Cloudera Altus?

Answer: Cloudera Altus provide almost the same solution which is provided by Cloudera Enterprise. But in the public cloud like AWS, Azure and Google Cloud.

Question-3: Can you please explain what is the use of Cloudera (SDX) Shared Data Experience?

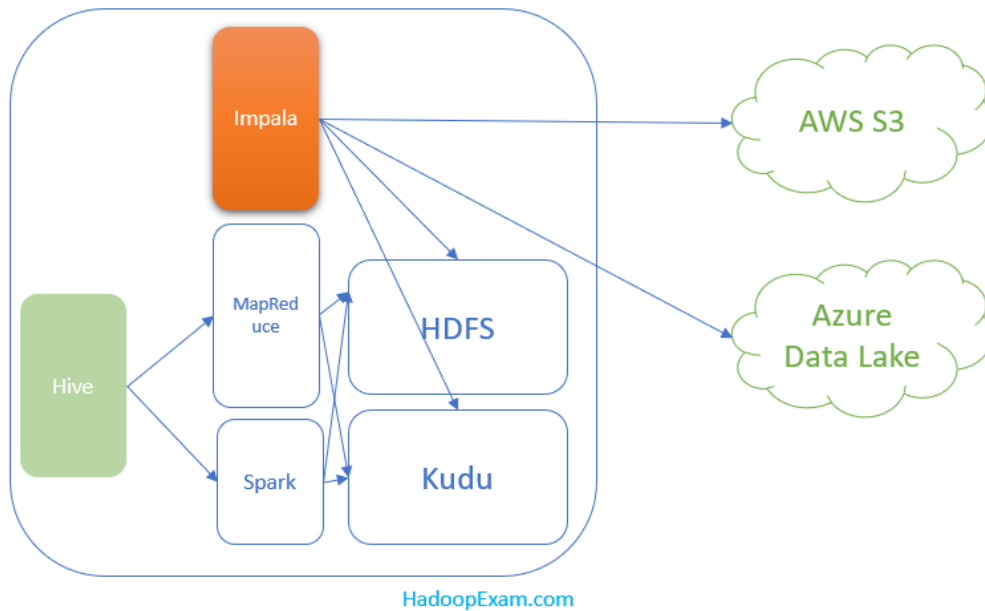
Answer: Using the Cloudera 's various solutions like Cloudera Enterprise we can have Data warehouse, data engineering, operational databases workloads altogether on the single platform. In such cases Cloudera Shared Data Experience (SDX) enables these diverse analytic processes to operate against a shared data catalog while having security, governance policies and schema. Even your entire Cloud environment is terminated, it still persists the all the metadata information.



Question-4: Can you please tell me which all components are used as part of Cloudera Data Warehouse solution?

Answer: Currently below 5 major components used to make Cloudera Data Warehouse solution.

- **Apache Impala:** You can run SQL/BI analytics on the data stored in either of the following
 - AWS S3
 - Microsoft Azure Data Lake
 - HDFS
 - Apache Kudu
- **Hive on Spark:** This helps in creating faster ETL/ELT solution for BI and Reporting.
- **HUE (SQL Development Workbench):** At a time 1000's of SQL developer.
- **Workload XM:** It is used to analyze the current workload, query analysis and optimization of the cluster resources.
- **Cloudera Navigator**



Question-5: As part of Cloudera Enterprise Data Science solution, which all are underlined product majorly used or it runs on?

Answer: Currently below 3 major components are used

- **Cloudera Data Science Workbench:** CDSW provides the on-demand access to Runtime for R, Python, Scala and integration with the Spark framework on CDH. Even for deep learning it supports the GPU accelerated computing and data scientists can use a framework like TensorFlow, MXNet, Keras etc.
- **Apache Spark:** Using Spark you can run in-memory processing.
- **Cloudera Fast Forward Labs:** Using this you can design and execute your enterprise machine learning strategy.

Question-6: What is the use of Apache Kudu?

Answer: Kudu is a Hadoop-native storage for fast analytics on fast data. It complements the capabilities of HDFS and HBase.

Question-7: What is Cloudera CDH?

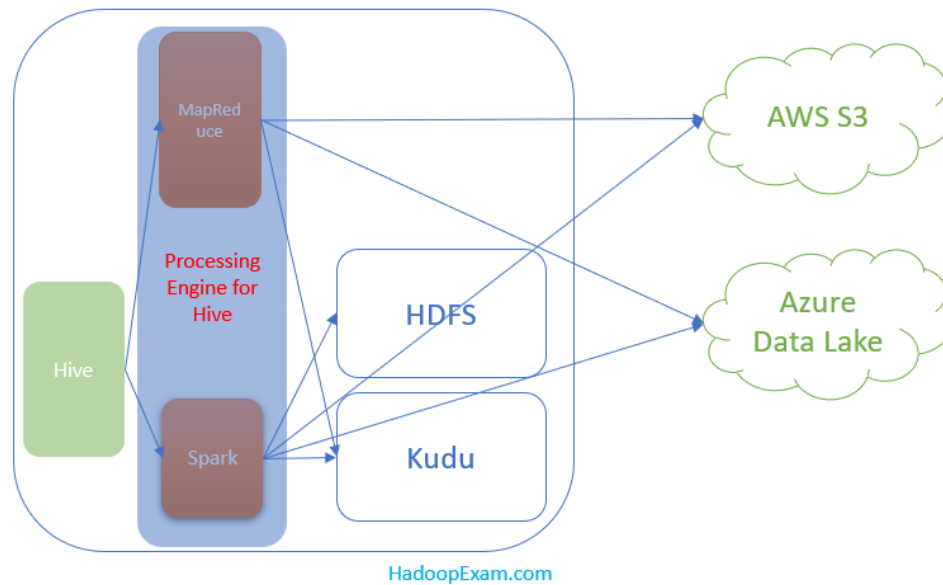
Answer: It is a distribution from Cloudera for Hadoop and its related projects. CDH is an open source product which include many projects few examples are below.

- Hive
- Impala
- Kudu
- Sentry
- Spark

CDH is considered unified solution for the batch processing, Interactive SQL, interactive search, Machine Learning, statistical computation and role-based access control.

Question-8: Please tell me something about the Apache Hive?

Answer: Hive is a data warehouse solution for reading, writing and managing large datasets in distributed storage like HDFS using Hive Query Language (Almost same as SQL). These queries are converted into a series of jobs which execute on a Hadoop Cluster using either MapReduce or Spark.



Question-9: There are many tools available for querying the data, then why to use Hive?

Answer: Hive is a petabyte-scale data warehouse system which is built on the Hadoop platform. And one of the best available choices where you expect high growth of data volume. Hive on either MapReduce or Spark is best suited for batch data preparation or ETL.

Question-10: Can you please give me some use cases where Hive should be used?

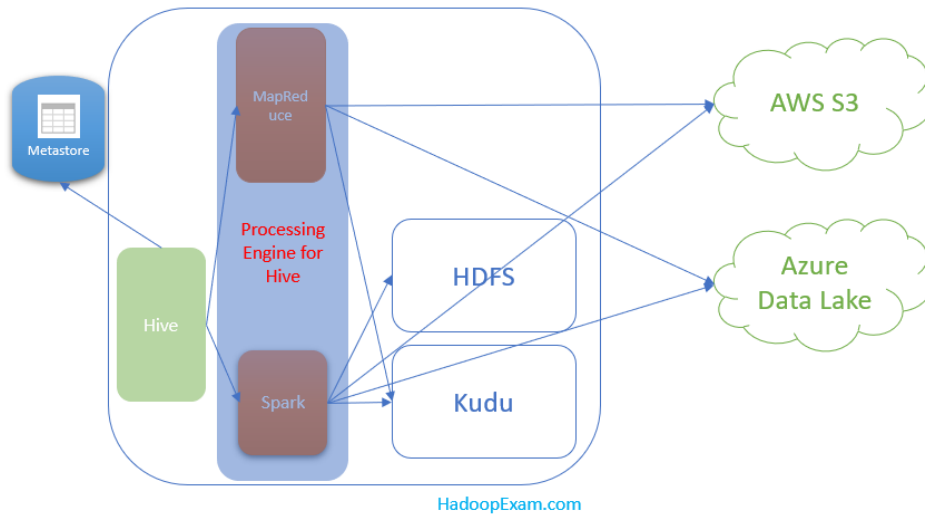
Answer: Let's see few of the below of the use cases

- Suppose you have large ETL Sort and Join jobs to prepare the data for BI users in Impala then schedule such ETL jobs in the Hive.
- Suppose you have a Job where data transfer or conversion take many hours and possibility of job failure in between then do such activity using Hive, which can help you in recovering and continues where it left.
- Various formats of the data, suppose you are receiving data in various formats then with the Hive SerDe and Variety of UDFs can help in converting data in single format.

Question-11: What is Hive Metastore?

Answer: Metastore is one of the RDBMS, which is required for Hive to work. It could be MySQL, PostgreSQL, Oracle etc. Usually metastore has following information (metadata) stored

- Name of the tables
- Columns in the table
- Partition information
- Hadoop specific information e.g. Data Files and their block locations.



Question-12: Can Hive metastore used by other Hadoop components?

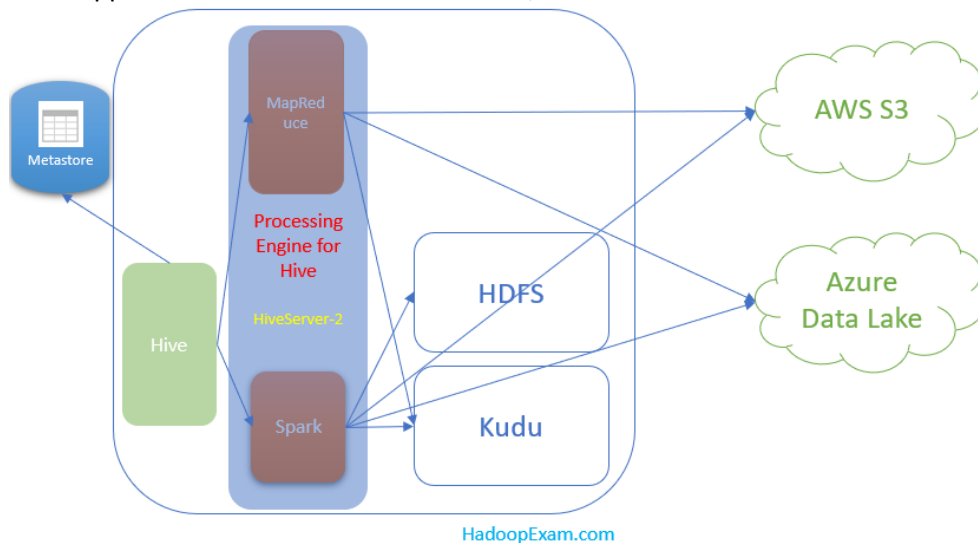
Answer: Yes, Hive metastore contains the information regarding data stored on HDFS, so that other Hadoop components like Impala can leverage that. Even if you don't have Hive then also this Metastore would be used.

Question-13: What do you mean by Remote Mode of Metastore?

Answer: Remote mode means metastore should be running in its separate JVM process. And any other process which wanted to get connected with the Metastore for example HiveServer2, HCatalog, Impala etc. should use the Thrift network API.

Question-14: What is HiveServer2?

Answer: HiveServer2 is a server-side interface, you can assume it as a container for the Hive Execution Engine. For each client connection it creates a new execution context for Hive SQL request submitted by the client. Hive support for both JDBC and ODBC client, which uses the Thrift API.



Question-15: Can Hive use the Apache Spark as a computation engine?

Answer: Yes, traditionally Hive using MapReduce as a computation engine, but Spark is much faster than MapReduce, hence in all modern solution Hive mostly uses the Spark as computation engine.

Question-16: Can you run Hive queries on HBase NoSQL database?

Answer: Yes. HBase is a NoSQL database which supports the real time read/write access to the large datasets in HDFS. We can run Hive queries on the HBase database as well.

Question-17: Can we use Amazon RDS (Relational Database Service) as Hive Metastore?

Answer: Yes, we can. Because AWS RDS is a service which provides the managed RDBMS solution like Oracle, MySQL etc.

Question-18: What is Impala?

Answer: Impala is fast query engine for running interactive queries on the data stored in HDFS, HBase or AWS S3. Even Impala use the same query syntax as Hive.

Question-19: Impala is a replacement for the Hive?

Answer: No. You can say Impala as an additional tool for querying BigData. Impala is better suited for interactive query, while Hive is better suited for the batch processing e.g. ETL.

Question-20: What are the advantages of the Impala over other existing BI/Reporting tools?

Answer: Following are the benefits of the Impala

- SQL query syntax same as existing SQL
- It can query high volume of the data on Hadoop.
- Distributed queries for high performance.
- Best fit with Hive. Because Impala can read from and write to Hive tables.

Question-21: What are all the possible clients for Impala?

Answer: You can have the following components as an Impala client, which can query or administer the Impala environment.

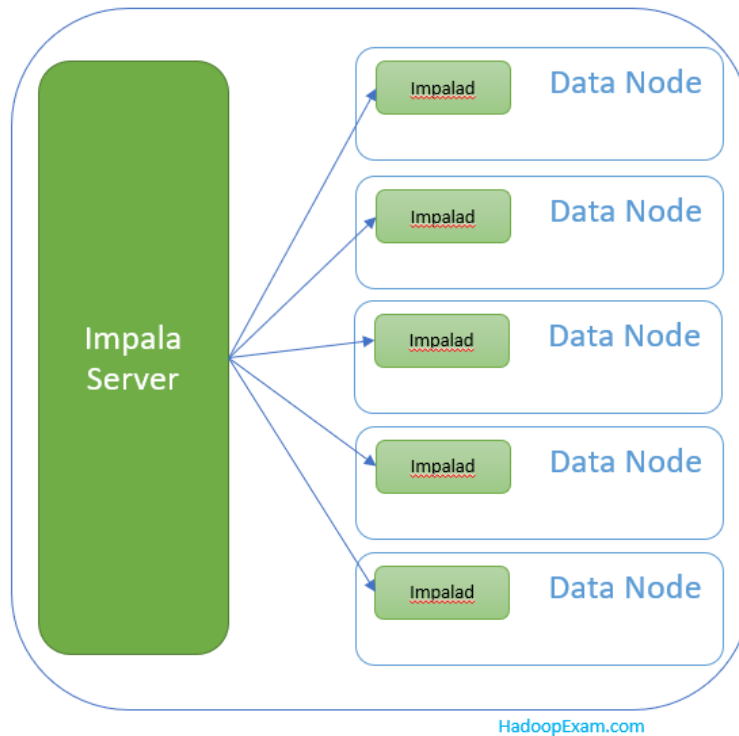
- Hue (Web Interface for querying)
- Impala Shell
- ODBC
- JDBC

Question-22: Can Impala use the Hive Metastore?

Answer: Yes, Hive Metastore has the information about available data and let it know structure of the data, schema, table name, column names etc.

Question-23: Can you please give me basic overview, how the queries are executed in case of Impala?

Answer: There is a process named Impala which runs on each DataNode on HDFS, which is responsible for executing and co-ordinating the queries. Each instance of the Impala can receive, plan & co-ordinate queries from Impala client. Queries would be distributed among Impala nodes, and these nodes then act as workers, execute queries in parallel.



Question-24: What is Apache Kudu?

Answer: Apache Kudu is a columnar storage manager, developed for Hadoop platform. Kudu also shares the same common technical properties of Hadoop Ecosystem as below

- Runs on commodity Hardware
- Horizontally Scalable
- Highly available operations

Question-25: Can you please tell me some benefits of the Apache Kudu?

Answer: Following are the few benefits of the Kudu

- Fast processing of OLAP workloads
- It can be easily integrated with the MapReduce, Spark, Flume & Other Hadoop Components.
- Tight integration with Impala
- Strong but flexible consistency e.g. consistency per request basis.
- Highly performant for running sequential and random workloads simultaneously.
- Can be managed using Cloudera Manager
- Structured Data Model
- Highly available

Question-26: What kind of applications where Kudu best fit?

Answer: There are following things which are difficult to implement on currently available Hadoop Technologies, but Kudu can help

- **Reporting application:** Where new data must be immediately available for end users.
- **Time-series applications:** Querying large amount of historic data as well as granular queries on individual entity.
- **Predictive Models:** Application which uses the predictive models for making real-time decisions, with the periodic refreshes of the predictive models based on historical data.

Question-27: What is Apache Sentry?

Answer: Apache Sentry is a granular, role-based authorization module for Hadoop. It is used as a plugin for authorization engine for Hadoop components. Using this we can define authorization rules to validate a user or application's access requests for Hadoop resources.

Question-28: On Cloudera CDH6, which all are cluster manager supported for Apache Spark?

Answer: Since CDH6 Spark Standalone Cluster Manager is no more supported, you have to use YARN as a cluster manager. On CDH6, Spark 1.6 is also not supported.

Question-29: What is the use of Cloudera Manager component?

Answer: Cloudera Manager is an end-to-end application for managing CDH clusters. With this we can easily deploy and centrally operate the complete CDH stack and other managed services.

Question-30: Can you please explain how Cloudera Manager, hosts and Cluster are associated?

Answer: Cloudera Manager is a logical entity that contains a set of hosts.

- On the hosts only single version of CDH is installed e.g. either CDH-5 or CDH-6
- Services and Role instances run on the Hosts e.g. HDFS instance
- A single host can belong to only one cluster.
- Cloudera Manager can manage more than one cluster.
- A single cluster can only be associated with the single Cloudera Manager

Question-31: Can you use virtual machine as a host in the Cloudera Manager?

Answer: Yes, in Cloudera Manager you can have either Physical or virtual machine as a host which runs the role instances. A host can belong to only one cluster.

Question-32: Can you tell me something about Rack?

Answer: Rack is a physical entity which contains a set of physical hosts typically served by the same switch.

Question-33: What do you mean by service or service type, with respect to Cloudera Manager?

Answer: In case of Cloudera Manager, you can say it is a category with the managed functionality which may or may not be distributed. Which is referred as either Service or service type. For example, MapReduce, HDFS, YARN, Spark, Accumulo etc. It is possible that same single service can run on more than one host.

Question-34: What is service instance?

Answer: You would be creating an instance from a service like HDFS is one service, you would be creating its instance by giving some name like "HDFS-HE" or "HDFS-A" etc. It is possible that a single service instance span across multiple Role instances.

Question-35: What do you mean by a Role in a service?

Answer: As you know a single service can have various functionality like HDFS has below functionality


- NameNode
- Secondary NameNode
- DataNode
- Balancer

Hence, role represent any of this functionality, sometimes this is also referred as Role Type. And similarly, for each role you would be creating a role instance. Each Role instance equivalent to a Unix process. For example, running a NodeManager instance.

Cloudera Certification & Training Material Available on HadoopExam.com are below. Please keep visiting website for that latest products

 95 Q & A Click Here	 73 Q & A Click Here	 90+ Solved Scenarios Click Here	 79 Q & A Click Here
---	---	---	--

Cloudera Hadoop & Spark Developer CCA175	Cloudera Hadoop BigData Analytics CCA159	Cloudera Hadoop Administrator Certification CCA131	Cloudera Data Engineer CCP:DE575
--	--	--	--

 Click Here	 57 Solved Scenarios Click Here
---	--

Package Deal	Hortonworks HDPCA : Hadoop Admin Certification
---------------------	---

Question-36: What is the Role Group?

Answer: Role group is a set of configuration properties for a set of role instances e.g. NodeManager-1, DataNode-1, Balancer-1 is one set and another set could be NodeManager-HE, DataNode-HE, Balancer-HE. Hence, we can create two Role groups something like that GROUP-1, Group-HE etc. to combine all the configuration properties for the same set of role instances.

Question-37: What is the use of Gateway node?

Answer: Gateway is again another type of Role, which give access to client for any specific service like accessing HDFS files, Running Hive Queries etc. on the cluster. Sometimes Gateway node are referred as Gateway Node or edge Node. It is mostly the node which is outside of the cluster having access to the Cluster service.

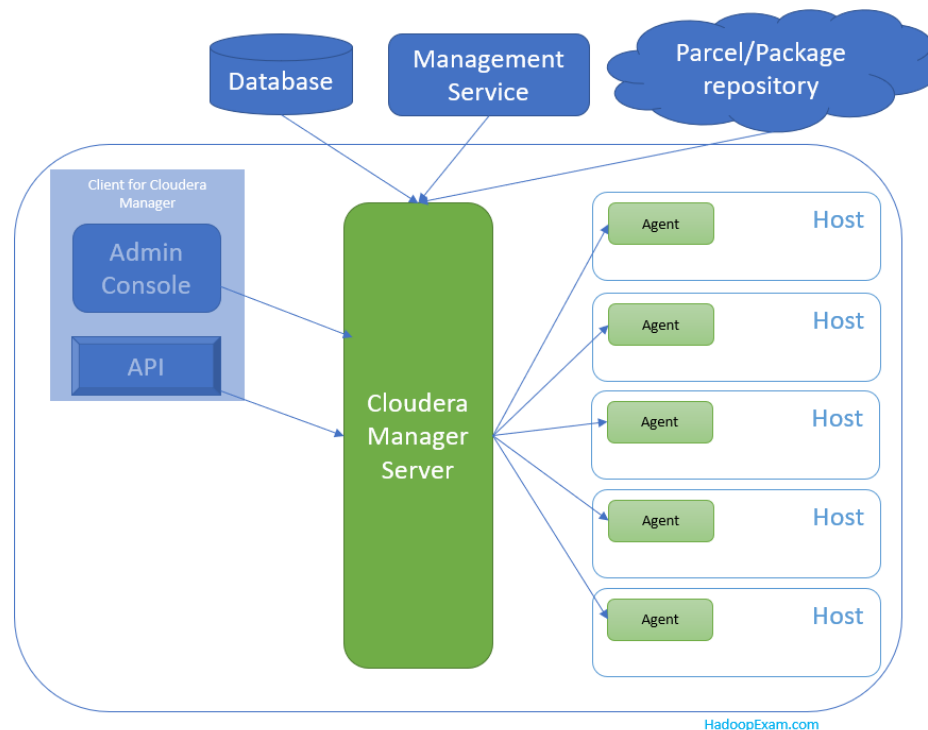
Question-38: What do you mean by Static Service Pool?

Answer: In Cloudera Manager, a static partitioning of total cluster resources e.g. CPU, Memory and I/O weight across a set of services.

Question-39: What is the role of Cloudera Manager Server?

Answer: Cloudera Manager server is heart of Cloudera Manager. Cloudera Manager Server hosts the

- Web Server for Admin Console
- Application logic
- It is responsible for installing software
- Configuring, Starting & stopping services.
- Managing the cluster on which the services run.



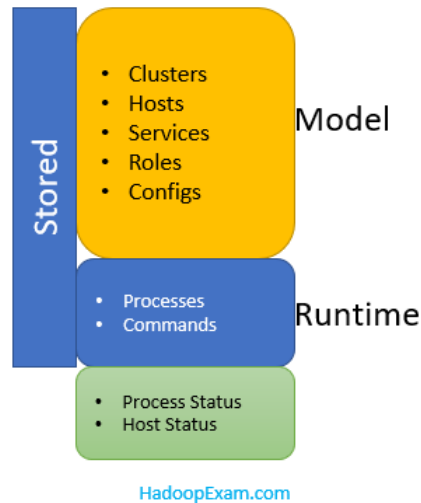
Question-40: How does Cloudera Manager able to start and stop services on so many nodes in Hadoop cluster?

Answer: There is an agent which needs to be installed on each host and this agent is responsible for starting and stopping various requested processes. Unpacking the configurations, triggering installations and monitoring the hosts.

Question-41: What are the two possible state of the Cloudera Hadoop Cluster as per the Cloudera Manager?

Answer: Cloudera Manager Maintains following two stats of the cluster

- **Model** - This tells the state of the Clusters, Hosts, Services, Roles & Config. Using the model state Cloudera Manager captures what is supposed to run where, and what configurations. Like 20 node cluster models can tell you that all 20 nodes should run DataNode.
 - **Runtime**: It tells what is currently and where it is running e.g. HDFS rebalancing
- Suppose you update any configuration like port for any service etc, you have updated the model state.



Question-42: Why do you see “outdated configuration” for one of the services?

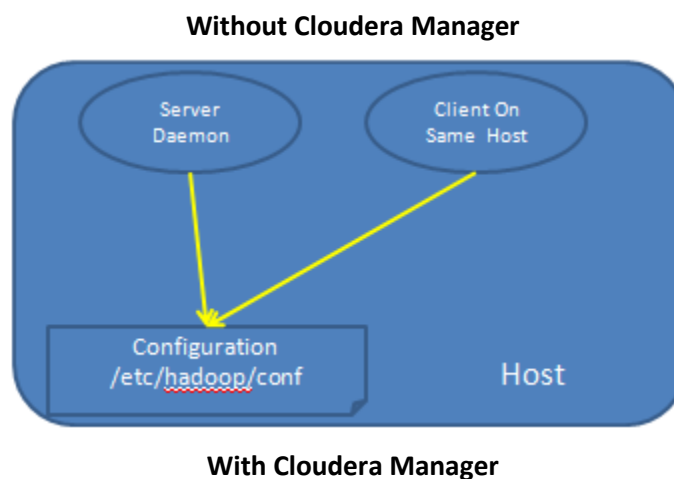
Answer: When you update the configuration let’s say Hue Web Server port, which update the model state. It still uses the old configuration like port. When such mismatch occurs then the role is marked with the “Outdated Configuration” and for synchronization we have to restart that role.

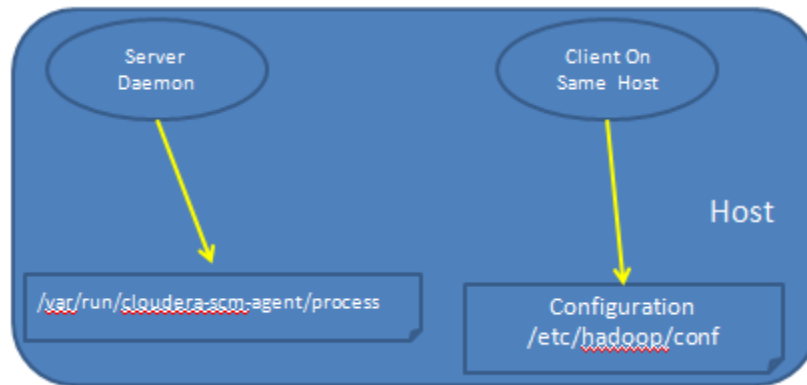
Question-43: What is Cloudera Express and Cloudera Enterprise?

Answer: Cloudera Express has the free version of Cloudera Manager, while Cloudera Enterprise does not have free Cloudera Manager, you can have 60 days trial version for that. And after that you need a license to use full featured Cloudera Manager.

Question-44: What do you mean that Cloudera Manager maintains the separate configuration for both Server and Client?

Answer: Service instance started by Cloudera Manager does not read configuration from the default location like /etc/hadoop/conf. If you are not using Cloudera Manager then both the Server-side daemon process and client running on the same host would read configuration on the same location e.g. /etc/hadoop/conf





However, Cloudera Manager distinguishes the Server and Client configurations. Client read the configuration from the default location as mentioned above while server-side Role instance read configuration from a private per-process directory under `/var/run/Cloudera -scm-agent/process/` This helps in not exposing sensitive information to the client like Passwords.

Question-45: With Cloudera Manager can you use init script to start and stop services?

Answer: No, Cloudera Manager does not use init scripts for the daemons it manages. Cloudera Manager uses the open source tool called “supervisord”, that starts processes, which takes care of redirecting log files, notifying of process failures, setting the effective user id of the calling process.

Question-46: If you stop Cloudera Manager then what happen?

Answer: Stopping the Cloudera Manager Server and Cloudera Agents will not bring down your services, any running role instances keep running.

Question-47: Which all are the software distributions are supported by Cloudera Manager?

Answer: Cloudera Manager support two software distribution formats

- **Packages:** Package is a binary distribution format which contains compiled code, meta-information, dependencies etc. Cloudera Manager uses the native system package manager for each supported OS.
- **Parcels:** Is a binary distribution format containing the program files, metadata etc.

Parcels are self-contained and installed in a versioned directory, which means that multiple versions of a given parcel can be installed side-by-side. You can then designate one of these installed versions as the active one. With packages, only one package can be installed at a time so there is no distinction between what is installed and what is active. If you want to have Rolling Upgrade enabled then parcels are also required and package does not support the rolling upgrades.

Question-48: What all things are managed by the YARN?

Answer: YARN manages the virtual cores, memory, running applications, maximum resources for pools, scheduling policy for each pool etc.

Question-49: What is Cloudera Navigator Data Management?

Answer: Cloudera Navigator Data Management is a complete solution for data governance, auditing, and other data management related tasks with the Hadoop platform, few of the example below.

- Finding what is the source of data?

- Is data altered, and if yes. Who altered it?
- Is data being used by downstream processes?
- Are unauthorized people trying to access the data?
- Data retention mandate is being met?
- Where is the most important data?

There are many such things can be Answered using the Cloudera Navigator.

Question-50: Is Cloudera Navigator is part of the Cloudera Manager?

Answer: No, Cloudera Navigator should be installed separately, Once the Cloudera Manager is installed. It interacts with the Cloudera Manager Behind the scene. As you already know Cloudera Manager is used to manage the cluster, while Cloudera Navigator is used by both administrator and also by security and governance team, data stewards, for auditing, tracing data lineage from source raw data through final form.

Question-51: What are the differences between the logs collected by Cloudera Manager and Cloudera Navigator?

Answer:

- **Cloudera Navigator:** Tracks and aggregates accesses to the data stored in CDH services and is used for audit report analysis.
- **Cloudera Manager:** It monitors and logs all activities performed by CDH services that helps administrators maintain the health of the cluster.

Question-52: Before installing Cloudera Manager, SELinux should be in which mode?

Answer: Cloudera supports the running of the Cloudera Software with SELinux enabled, However, Cloudera Manager installer would not proceed if SELinux is enabled. Hence, we need to disable SELinux or set it to permissive mode before running the installer. Once Cloudera Manager and CDH deployment is done, you can re-enable SELinux by changing SELinux=permissive back to SELINUX=enforcing

Question-53: Can you please tell me something about Upgrading CDH using Cloudera Manager?

Answer: We can upgrade the CDH within the Cloudera Manager Admin Console using parcels. Once CDH is upgraded using parcels, you can perform rolling upgrades on your CDH services. If HDFS high availability configured and enabled the we can do the rolling upgrades on the cluster without bringing down the entire cluster.

Question-54: Why i am not able to find the CDH libraries when CDH is distributed using parcels?

Answer: Because in case of parcels all the libraries are installed on different path at /opt/Cloudera /parcels/CDH/lib and not at under /usr/lib

Question-55: I don't have access to public internet, can I use the local repository to install software?

Answer: In both the cases whether you use parcels or packages it is possible to create local repositories that serves these files to hosts that are being upgraded. And no need to access Cloudera public repositories.

Question-56: Is it required by each Worker host to access repository to install the software?

Answer: No, If you are using the parcels, in this case on Cloudera Manager Server requires access to the Cloudera Public repositories. Distribution of the parcels to worker hosts is done between the Cloudera

Manager Server and the worker hosts. If you are using traditional packages then host only requires access to the installation files.

Question-57: Which feature is used, so that text SQL queries are not visible in the logs?

Answer: You can use the log redaction feature to obfuscate sensitive information in the impala log files.

Question-58: Is it required to install Impala on all the nodes in the cluster?

Answer: Yes, it is important to install Impala on all the DataNodes in the cluster. Because otherwise some of the nodes must do remote reads to retrieve data not available locally. Because data locality is an important aspect of the Impala performance. As the number of nodes increases Impala performance also increases.

Question-59: During the query processing by Impala to improve the query performance HDFS block size is reduced?

Answer: No, Impala does not change the block size of HDFS and not even it changes HBase dataset size.

Question-60: Is Impala uses the caching for faster query result?

Answer: No, Impala explicitly does not cache the data. It caches some of the metadata for the files and tables.

Question-61: Impala does not cache the data; then why subsequent run of the query is faster?

Answer: Impala does not cache the data but subsequent run of the query is faster because the data set was cached in the OS buffer cache, Impala does not control this explicitly.

Even, Impala takes advantages of the HDFS caching feature in CDH. Like while creating table we can designate which tables or partitions are cached explicitly through CACHED and UNCACHED clause.

HDFS Cache: Impala can also take advantage of data that is pinned in the HDFS cache through the hdfs-cache-admin command.

Question-62: Where do you prefer Impala instead of Hive or MapReduce?

Answer: Impala is well suited for executing SQL queries for interactive exploratory analytics on large datasets. Hive and MapReduce are appropriate for very long running, batch-oriented tasks such as ETL.

Question-63: Impala Uses the MapReduce as an underline processing engine?

Answer: No, Impala does not use the MapReduce. Even you stop MapReduce service Impala would work fine.

Question-64: Can we use the Impala for Stream Processing?

Answer: Stream-processing or Complex Event processing is not well suited for the Impala. Because it is most closely resembling a relational database.

Question-65: How you compare Impala with Hive and Pig?

Answer: Impala is different than Pig and Hive, because it uses its own daemons that are spread across the cluster for queries. Impala does not use the MapReduce where Pig and Hive do, and not using the MapReduce avoid the startup overhead and allowing Impala to return the results in real-time.

Question-66: How Impala query and Hive Query are related?

Answer: There are some minor differences between Impala query and Hive Query. However, Impala queries can be executed in the Hive, because Impala SQL is a subset of HiveQL.

Question-67: How does it affect Impala Query if data is already loaded in the HBase or Hive?

Answer: It does not matter, the only requirement is that Impala should be able to access Hive metastore. Keep in mind that impalad, by default, runs as the Impala user, so you might need to adjust some file permissions depending on how strict your permission are currently.

Question-68: Is Hive required to run the Impala?

Answer: Hive metastore is required by the Impala, because Impala shares the same metastore database as Hive, allowing Hive and Impala to access the same tables transparently. Hive itself is optional and does not need to be installed on the same nodes as Impala. As Impala has more variety of read query instead of write. And Hive provides more option to insert data in the table.

Question-69: Can Impala able to query the table which has trillions of rows?

Answer: Yes, many of the Cloudera Customer achieved this.

Question-70: Can I configure Impala for High Availability?

Answer: Yes, you need to set up a proxy server to relay requests back and forth to the Impala servers, for load balancing and high availability. For Hive metastore enable the HDFS HA.

Question-71: Is Impala Single point of failure?

Answer: No, all impala daemons are fully able to handle incoming queries, if a machine fails, all queries with fragments running on that machine will fail. Because queries are expected to return quickly, if there is any failure then you can re-run query.

Question-72: How Impala, Hive, HDFS HA, NameNode are linked or related?

Answer: Impala uses the same Hive Metastore, and aggressively caches the metadata so the metastore host should have minimal load. Impala relies on the HDFS Namenode, and you can configure HA for HDFS. Impala has centralized services known as the statestore and catalog services, that run on one host only. Even if statestore host is down Impala continues to execute queries and would not get state updates.

Question-73: What happen Impala statestore is down?

Answer: Suppose a new host is added in the cluster when statestore is down then the existing instances of the impalad running on the other hosts will not find out about this new host. Once the statestore process is restarted, then all the information it serves is automatically reconstructed from all running Impala daemons.

Question-74: Why it is recommended that Impala daemon should run on each DataNode?







Answer: It is highly recommended that impalad daemon should be running on each DataNode in the cluster to avoid any kind of remote data read and affecting the query performance. If possible, impala schedules query fragments on all hosts holding data relevant to the query.

Question-75: Between Small and large table how joins are performed?

Answer: There are various strategies based on the size of the table's joins are performed. When two tables are joined where one is a large table and another is small table then data from the small table would be transmitted to each node for intermediate processing. This is also known as broadcast join.

When both the tables are large than data from one table is divided into pieces, and each node processes only selected pieces.

Cloudera Certification & Training Material Available on HadoopExam.com are below. Please keep visiting website for that latest products

 95 Q & A Click Here	 73 Q & A Click Here	 90+ Solved Scenarios Click Here	 79 Q & A Click Here
Cloudera Hadoop & Spark Developer CCA175	Cloudera Hadoop BigData Analytics CCA159	Cloudera Hadoop Administrator Certification CCA131	Cloudera Data Engineer CCP:DE575
 Click Here	 57 Solved Scenarios Click Here		
Package Deal	Hortonworks HDPCA : Hadoop Admin Certification		

Question-76: During the aggregation sometime Impala uses the disk space as well, why?

Answer: Because Impala currently supports only In-memory hash aggregation. In Impala 2.0 onwards if the memory requirements for join or aggregation operation exceed the available memory limit on a particular host then it uses the work area on the disk to help the query for completing successfully.

Question-77: Which all metadata are used by Impala currently?

Answer: There are two types of metadata is being used currently

- Catalog information from Hive Metastore
- File Metadata from the NameNode

Both these metadata lazily populated (means whenever they are needed) and then cached. Using the REFRESH statement, we can update the metadata for a particular table. Using the INVALIDATE METADATA statement all metadata are refreshed. Hence, Impala can recognize the new tables or DDL and DML changed done through Hive. In Impala 1.2 or later these statements are not needed because a daemon named catalogd broadcasts metadata changes to Impala.

Question-78: For what Impala uses the NameNode?

Answer: Impala connects with the NameNode during the planning phase to get the file metadata to send the query on the host which has the data. Every impalad will read files as part of normal query processing.

Question-79: With the design perspective, can you tell me why Impala is considered faster query engine?

Answer: There are many reasons because of that Impala is faster than other Hadoop components.

- Impala does not use the MapReduce because MapReduce has few processing inefficiencies.
- Impala does not materialize intermediate results to disk.
- Impala does not have start-up time because it does not use the MapReduce.
- Impala runs as a service and essentially does not have start-up time.
- Impala does not create a pipeline of Map & Reduce job to run the query, rather disperse query plans. And avoid all the overheads of sort and shuffle phase when not needed.

Question-80: What all Hardware feature used by Impala?

Answer: Impala uses the more efficient execution engine by taking advantage of modern Hardware and technologies,

- Impala generates runtime code and uses the LLVM to generate assembly code for the query that is being run.
- Whenever, possible Impala uses the available hardware instructions.
- Impala uses better I/O scheduling, because it is aware about the location of the data block by reading the metadata from NameNode.
- Many other things have been taken care as below
 - Tight inner-loops
 - Inline function calls
 - Minimum branching
 - Better use of cache
 - Minimal memory usage

Question-81: What happens to query which has more data size then available memory?

Answer: As of now, if the memory required to process intermediate result on a node exceed the available memory to impala process then the query would be cancelled. We can even adjust the available memory on each individual node and fine tune the strategy. Because currently external join and sorting is not supported for Impala.

Question-82: Why do I see higher memory usage by Impala, even query is not running?

Answer: Impala allocates memory and once allocated it keeps this memory reserved for future use, the name of the memory allocator is tcmalloc (optimized for high concurrency). Hence, if you are a programmer and using JDBC/ODBC than call appropriate close method afterwards. Otherwise, some memory associated with the query will not be freed.

Question-83: Impala supports UDF?

Answer: Yes, you can use UDFs and UDA and need to be written using C++, and existing UDFs from the Hive can also be used.

Question-84: When a managed table is dropped, still disk space is not freed, why?

Answer: When you drop a managed table, it moves the data in another trash directory, hence disk space is not free for some configured like 6 hrs.

Question-85: What kind of data best fit for HBase database?

Answer: HBase are good where you need to store key-value data and query needs to fetch few rows from the table, using = or IN operator.

And you should avoid HBase if your query needs to fetch rows more that few thousands. Full table scan using where clause is worst for the HBase table. Often HBase tables are wide and sparse many of the values in a row may be Null.

Question-86: Which all are the filesystem supported by HDFS as of now?

Answer: Following file systems are recommended as of now

- Ext3
- Ext4
- XFS: This is default on the RHEL7
- AWS S3

Question-87: Why it is said that on Linux system file read operations also leads to a write operation?

Answer: Linux filesystem keep metadata that records when each file was accessed. This means that even reads results in a write to the disk. Hence, to speed up the file reads, it is recommended that you disable this feature.

Question-88: Why it is recommended that during the filesystem mount, we should not use the sync option?

Answer: The filesystem mount options have a sync option that allows you to write synchronously. Using the sync filesystem mount option reduces the performance for services that write data to disks, such as HDFS, YARN, Kafka, and Kudu. Because most of the time write operation is already replicated and synchronous writes to disk is unnecessary, expensive, and do not measurably improve stability.

Question-89: How do you explain the High Availability and Load Balancing?

Answer: Load balancing refers to distributes the operations across multiple services let's say RDBMS instances in parallel. While HA focuses on the service continuity. However, Load balancing deployment often used as part of HA strategies to overcome demands and monitoring and failover management in HA environment.

However, in the case of Cloudera it is clearly mentioned that components are not designed to support load balancing. During HA strategy with multiple instances ensure that connections are routed to a single RDBMS service at any given time.

Question-90: What do you mean by ordinary objects pointers In Java?

Answer: This is an Optimization technique in Java Which is also known as Compressed oops Which enable 64-bit JVM to address heap size up to 32GB using 4-byte pointers. For large heap size 8-byte pointer are required. This means that heap size slightly less than 32GB can hold more objects than a slightly more than 32GB.

So, it is recommended that if you don't need heap size more than 32GB then use heap size 31GB or less to avoid this issue. if you need 32GB or more then that your heap size to 48 GB or higher to account for the larger pointers, so whenever you need heap size more than 32GB then multiply the amount of heap you need by 1.5

Question-91: Is CDH support IPV6 Protocol?

Answer: no, CDH require ipv4 protocol. and you should disable the IPV6 Protocol

Question-92: Why Cloudera manager agent runs using root user?

Answer: Cloudera manager agent runs as a root user so that it can make sure that the required directories are created and that processes and files are owned by the appropriate user for example hdfs and mapred user.

Question-93: is it true that firewall should be disabled to install CDH cluster?

Answer: yes, Firewalls such as iptables and firewalld must be disabled or configured to allow access to ports used by Cloudera Manager, CDH or other services.

Question-94: In CDH cluster at what all are the levels within the hadoop data at rest should be encrypted?

Answer: data at rest encryption protection can be applied at a number of levels within the Hadoop as below

- OS Filesystem level
- Network Level
- HDFS level

Question-95: Using Cloudera manager with various Browser should we enable cookies or JavaScript?

Answer: Yes, Cookies and JavaScript must be enabled.

Question-96: Which all cloud platforms are supported by Cloudera manager?

Answer: Currently there are three cloud platform are supported as below

- Amazon Web service
- Google cloud platform
- Microsoft Azure

Question- 97: Which service is used for time synchronization across all the host in the cluster?

Answer: CDH requires that you configure a network time protocol (NTP) service on each machine in your cluster.

Question-98: is Hue require python?

Answer: yes, with CDH 6, Python 2.7 is required.

Question-99: when you use parcels or packages, what repository accesses are required?

Answer: for a parcel installation, only the Cloudera manager server need internet access, but for a package installation all Cloudera host require access to the Cloudera repository.

Question-100: Can you please explain the parcels and why it does not require internet connection for installation?

Answer: Parcels are packaging format that facilitate upgrading software from within Cloudera manager. you can download, distribute, and activate a new software version all from within Cloudera manager. Cloudera manager download parcel to a local directory. Once the parcel is downloaded to the Cloudera manager server host, an internet connection is no longer needed to deploy the parcel. Even if your Cloudera manager server does not have internet access, you can obtain the required parcel files and put them into a parcel repository.

Question- 101: What all steps you have to do for using internal remote parcel repository?

Answer: if you want to have internal remote parcel repository then you need to follow these three things

- Set up a web server to host internal repository.
- Download and publish the parcel repository.
- Configure Cloudera manager to use an internal remote parcel repository

Question-102: Can you give an example where you have to manually install Cloudera software packages?

Answer: installing Cloudera software packages manually like Cloudera manager and cdh, is useful for the environment where it is not possible to use Cloudera manager to install the required packages, such organization where password less Sudo is not permitted. However clouded does not support the cluster which is not deployed or managed by Cloudera manager.

Even there are some services which is available only through parcel, cannot be installed using manual packages.

Question-103: Can you create virtual images for the Cloudera cluster host?

Answer: Yes, you can create virtual machine images, like PXE- boot images, Amazon AMIs, Azure VM image of cluster host with pre-deployed Cloudera software. Once this image has been created, you can use it to quickly spin up virtual machines. These images use parcels to install CDH software.

Question-104: Which location Cloudera recommend to install JDK?

Answer: Cloudera strongly recommend installing the JDK at /usr/java/jdk-version, which allows Cloudera manager to auto-detect and use the correct JDK version. if you install the JDK anywhere else, then you have to configure Cloudera manager with the chosen location.

Question-105: What all are the general steps followed to setup Cloudera CDH cluster?

Answer: generally, you would be following the steps below in order to install Cloudera CDH cluster.

- Configure the repository for parcels or packages.
- Then install the JDK
- Install Cloudera manager server
- Install required data bases
- Set up the Cloudera manager database
- Install CDH and other software
- Finally, setup a CDH cluster

Question-106: Can you have different version of the JDK across the host in CDH cluster?

Answer: No, you can use different JDK like Oracle provided JDK, or open JDK. But same version of JDK must be installed on each cluster host. As mentioned previously, it is highly recommended that you install the JDK at `/usr/java/jdk-version`

Question-107: What happens when you use G1GC garbage collector of Java?

Answer: When using G1GC garbage collection, then pauses for garbage collection are shorter, component will usually be more responsive, but they are more sensitive to JVMs with over committed memory usage.

Question-108: While installing Cloudera manager server, why it is recommended to enable AUTO-TLS?

Answer: Because auto TLS simplify the process of enabling and managing TLS encryption on your cluster. It automates the creation of an internal certificate authority and deployment of certificate across all cluster host. It can also automate the distribution of existing certificate, such as those signed by a public certificate authority. So whenever you add a new host or service to your cluster with auto TLS enabled, it automatically create and deploy the required certificates.

Question-109: What all are the recommendation by the Cloudera for the databases?

Answer: Cloudera manager uses various data bases and data store to store information about Cloudera manager configuration, as well as information such as the health of the system or task progress.

Following are the recommendation from Cloudera

- Choose one of the supported database providers for all of the Cloudera databases.
- It is recommended that install the database is on different hosts then the services, which helps in isolating the potential impacts from failure or resource contention in one or the other.

Question-110: What information is stored by Cloudera manager server database, and how it helps during restart?

Answer: Cloudera manager server database contains all the information about services you have configured and their role assignments, all configuration history, commands, users, and running processes. This is relatively small database < 100MB, but it is important to create backup. Because when you restart processes, the configuration for each of the service is re-deployed using information saved in the Cloudera manager database. If this information is not available, your cluster cannot start functioning correctly. Has it is highly recommended that you schedule and maintain regular backups of the Cloudera manager database to recover the cluster in the event of the loss of this databases.

Question-111: What is the different type of cluster you can create?

Answer: you can create following two type of cluster

- **Regular Cluster:** A regular cluster contains storage nodes, compute nodes, and other services such as metadata and security co-located in a single cluster.
- **Compute cluster:** A compute cluster consists of only compute nodes, connect to an existing storage, metadata for security services, you must first choose or create a data context on a base cluster.

If you are doing new installations, Regular Cluster is the only available option. You cannot have a compute cluster if you do not have an existing base cluster.

Question-112: What exactly Cloudera Enterprise?

Answer: Cloudera Enterprise brings open source Hadoop framework and its Commercial support to the platform.

Question-113: In Cloudera enterprise where does Cloudera navigator fit in?

Answer: It helps in security with the following functionalities

- **Navigator Encrypt** also known as NavEncrypt, which encrypt and secure data at rest.
- **Navigator Key Trustee Server:** Which works as key management store, which is used by Navigator Encrypt, which separates the encryption keys from your data.
- **Navigator Key HSM:** This is a Hardware security module, which provides the highest level of security for your encryption keys.
- **Navigator Audit Server:** Collects audit events from cluster services and provide central searchable audit dashboard.
- **Navigator Metadata server:** collects technical metadata from assets in your cluster and uses that information is used to show how this data is being used and how it changes over time.

Both Audit server and Metadata server share the single user interface, which is Navigator console. Combining this Navigator products with the access policy and enforcement from Apache Sentry and the security management features of Cloudera Manager, provides comprehensive security and governance. Hence, we can say that Cloudera Navigator offers the components for security, data management, and data optimization.

Question-114: Cloudera Navigator divided in three major parts what all are those?

Answer: These three are below

- **Cloudera Navigator Data Encryption:** This is a component used for data at rest encryption and also include Key Management suite, NavEncrypt and Key Trustee Server.
- **Cloudera Navigator Data Management:** is a comprehensive auditing, data governance, compliance, data stewardship, and data lineage discovery component that is fully integrated with Hadoop.
- **Cloudera Navigator Optimizer:** This is an analysis tool that profiles and analyzes query text in SQL workloads, which is used to better understand the workloads, identify and offload queries best suited for Hadoop. Even it optimizes the workloads which is already running on Hive and Impala.

Question-115: Can Cloudera navigator be installed without Cloudera manager?

Answer: No, Cloudera Navigator auditing and meta data subsystem interact with various Cloudera manager subsystem. Cloudera manager is a prerequisite for Cloudera navigator and must be installed before you can install Cloudera navigator.

Question-116: What is the difference between Cloudera navigator logs and Cloudera manager logs?

Answer: Navigator tracks and aggregate access to the data stored in CDH services and it used for audit report and analysis. Cloudera manager monitors and logs all the activity performed by CDH services that helps administrators maintain the health of the cluster.

Question-117: Is Cloudera Navigator Key Trustee Server specific to host?

Answer: Yes, when the key Trustee server role is created it is tightly bound to the Identity of the host on which it is installed. Moving the role to a different host, changing the hostname, or changing the IP of the host is not supported. You can install navigator key Trustee server using Cloudera manager with parcels.

Question-118: What is AES-NI?

Answer: The AES-NI is known as Advanced Encryption Standard New Instructions; this is an instruction set. That is designed to improve the speed of encryption and decryption using AES. Some newer processors come with AES-NI, which can be enabled on per server basis.

Question-119: What is navigator key HSM?

Answer: Cloudera navigator key HSM is a Universal hardware security module driver that translates between the target HSM platform and Cloudera navigator key Trustee server. With the HSM you can use a key Trustee server to securely store and retrieve encryption keys and other secure objects, without being limited solely on hardware-based platform. Keep in mind that you install key HSM on the same host as the key Trustee server.

Question-120: What is the key Trustee KMS?

Answer: Key Trustee KMS is a custom key management server (KMS) that uses Cloudera navigator key Trustee server, which is the underline key Store, instead of the file-based Java KeyStore (JKS) which is used by the default Hadoop KMS.

Question-121: Previously you had to set up a Cloudera Hadoop (CDH) cluster using package, can you migrate to parcels?

Answer: Yes, it can be done, managing software distribution using parcel, which offer many advantages over the packages. However, reverse is also possible.

Question-122: What all things you do to secure your CDH cluster?

Answer: To make the cluster setup secure we have to enable authentication ,authorization, auditing and encryption.

Question-123: What all are the features of the Cloudera Manager?

Answer: Below are the list of features

- Cloudera manager is end to end application for managing CDH cluster.
- Cloudera manager we can easily deploy and centrally operate the complete CDH stack and other managed services.
- It automates the installation process, and reduce the deployment time.
- It provides cluster wide, real time view of host and services running.
- It provides single Central console to enact configuration changes across your cluster.
- Cloudera manager incorporates a full range of reporting and Diagnostic tools for Optimization and utilization.
- Even Cloudera manager provides an API so you can use to automate clustered operation.

Question-124: In Cloudera manager, what does it mean “client configuration redeployment required”?

Answer: It indicates that the client configuration for a service should be redeployed.

Question-125: What is the requirement if you want to move Cloudera manager server on different host?

Answer: We can move the Cloudera manager server if either the Cloudera manager database server or current backup of the Cloudera manager database is available.

Question-126: What is Cloudera Management Service?

Answer: Using Cloudera Management Service various management features are implemented for example below

- **Activity Monitor**-Collects information about activities run by the MapReduce service.
- **Host Monitor**-Collect health and metric information about the Host.
- **Service Monitor**-Collects health and metric information about services and activity information from the YARN and Impala services.
- **Event server**-Aggregates relevant Hadoop events and makes them available for alerting and searching.
- **Alerts**: Generates and delivers the alerts for certain types of events.
- **Report Manager**: Reports for Historical view.

Cloudera Management service manages each role separately, instead of as part of the Cloudera Manager server.

Question-127: In Cloudera Cluster, how service and roles are related?

Answer: Whenever you configure a service for the Cloudera Manager then it allocates the various possible roles like HDFS, this is one service and under which there are various roles like NameNode, SecondaryNameNode, DataNode etc.

- One host would run NameNode Role
- One host would run the Secondary NameNode role
- Another Host would run the Balancer Role
- Remaining Hosts run DataNode roles.

Question-128: What is Role Group?

Answer: A role group is a set of configuration properties for a role type, as well as of role instances associated with that Group. Cloudera Manager automatically creates a default role group named Role Type Default Group for each role type.

Question-129: What is Client Configuration Files?

Answer: To unlock client to use the HBase, Hive, MapReduce, YARN, MapReduce etc. Cloudera manager creates zip archive after configuration files containing the service properties, and this zip archive is referred as a client configuration file. Each archive contains the set of configuration files needed to access the service for example the MapReduce client configuration file contains copies of

- core-site.xml
- hadoop-env.sh
- hdfs-site.xml
- log4j.properties
- Mapred-site.xml

Client configuration files are generated automatically by Cloudera manager based on the services and Roles you have installed and Cloudera Manager deploys these configurations automatically when you install your cluster, add a service on a host or at a Gateway Role on a host.

Question-130: How the client configuration is deployed?

Answer: Client configuration files are deployed on any host that is a client for a service. For example, Data node client configuration installed on the same node where the data node role is installed. however, you can use this client configuration on another host as well. Cloudera manager will

deploy client configuration files automatically in many cases, if you have modified the configuration for a service, then you need to redeploy those configuration file.

Question-131: Why do you want to refresh the cluster?

Answer: We do cluster refresh action to bring configuration up to date without restarting all services.

Question-132: When can I pause CDH cluster in AWS?

Answer: If all of the data for a cluster stored in EBS volumes, then You can pause the cluster and stop your EC2 instances during periods when the Cluster will not be used. When the cluster is paused it is not available and cannot be used to process the data, however it helps in reducing the cost of EC2 instances. Whatever storage you are using from EBS volumes would incur the cost even when cluster is passed. It is must to use EBS volume for your storage, whether it's management or Worker nodes. because data stored on an ephemeral disk will be lost as soon as EC2 instances are stopped.

Cloudera Altus

Question-133: What is Cloudera Altus Data Engineering?

Answer: Using Cloudera Altus Data Engineering you can create a cluster, with different kind of distributed processing engine like Spark, Hive, Hive on Spark, & MapReduce2(MR2). Once the cluster is created you can run data engineering and Data science, Machine Learning jobs on it.

Question-134: Where can i create cluster using Cloudera Altus?

Answer: Altus Data Engineering usage or access your AWS account or Azure subscription to create cluster in Cloud and run the job on the same cluster.

Question-135: What do you mean by Cross-Account access in case of AWS & Cloudera Altus?

Answer: If you are creating a Cluster using Cloudera Altus then AWS Administrator must setup a cross-account **access role** to provide altus access to your AWS account. If any of the Altus Data Engineering account holder create a Cluster in AWS that time Altus Data Engineering service uses the Altus Cross-Account access credentials to create the cluster in your AWS account.

Question-136: What is the requirement to create the Cloudera Altus Data Engineering cluster in the Microsoft Azure?

Answer: If you are using Azure subscription then your Administrator of the Azure subscription must provide the consent for Altus to access the resources in your subscription. If any of the account holder in the Altus create a cluster then Azure Administrator consent allows Altus to create the cluster under that Azure subscription.

Question-137: Should I keep running the Altus cluster?







Answer: No, Altus manages the cluster and jobs in your cloud provider account (AWS, Azure). If you don't need the cluster then you should configure Altus such a way that Cluster should be terminated when the cluster is not in use. So, that cost can be saved.

Question-138: Where does Altus input/output the data when it needs while running the job?

Answer: As soon as you submit the job which needs to be executed on the cluster, then Altus create a Job Queue then this job would be executed on the cluster. Cluster can be on AWS or Azure (Depend

where you have created). During the Job execution if data is needed then in case of AWS, S3 object storage would be used. In case of Azure it would be using Azure Data Lake Store. Even in that storage Altus also stores the cluster and Job information in your Cloud object storage.

Cloudera Certification & Training Material Available on HadoopExam.com are below. Please keep visiting website for that latest products

 95 Q & A Click Here	 73 Q & A Click Here	 90+ Solved Scenarios Click Here	 79 Q & A Click Here
Cloudera Hadoop & Spark Developer CCA175	Cloudera Hadoop BigData Analytics CCA159	Cloudera Hadoop Administrator Certification CCA131	Cloudera Data Engineer CCP:DE575
 Click Here	 57 Solved Scenarios Click Here		
Package Deal	Hortonworks HDPCA : Hadoop Admin Certification		

Question-139: Can you please explain Cloudera Altus Engineering Service once again?

Answer: Cloudera Altus is a service for the Public Cloud Platform as of now AWS and Azure. You can create CDH cluster and once done you can configure it to auto terminate the cluster. Altus helps to provision the CDH cluster quickly and make it very easy for you to build and run your data workloads in the Cloud.

You have to choose whether you want Azure or AWS, so that Altus can create CDH cluster either one of them.

Question-140: How does Altus uses networking in respective public Cloud for Cluster creation?

Answer: Altus uses the following networking solution

- AWS: Altus cluster would be created inside Virtual Private Network in your account.
- Azure: Under you Azure subscription, in vNet (Virtual Network) it would be created.

Question-141: What are all the interfaces available with the Cloudera Altus?

Answer: Currently there are following 3 interfaces available

- CLI: Command Line Interface

- SDK for Java
- Web User Interface

And using all above three we can

- Create and Manage environments
- Create and Manage Clusters
- Run Jobs
- Performing tasks of the Altus itself

Question-142: What are all the services available for Cloudera Altus?

Answer: There are mainly three services available as of now, which can be used for your different kind of workloads.

- Altus Data Engineering Service
- Altus Data Warehouse Service
- Altus Shared Data Experience (SDX) Service

Question-143: What is the use of Altus Data Engineering Service?

Answer: Using the Altus Data Engineering service you can create Cluster, run jobs which are related to Data Science and Data Engineering related. You can have any of the following distribution engine

- Hive
- Spark
- Hive on Spark
- MapReduce2(MR2)

And submit the jobs, which can be

- ETL
- Machine Learning
- Large Scale Data Processing

Question-144: What is the use of Altus Warehouse Service?

Answer: Altus Data Warehouse service helps you to create the Clusters which runs the Impala SQL engine to access data in your Cloud Storage for business analysis and reporting. There are following options available to access and analyze the data.

- Query editor: To query the data
- ODBC/JDBC connector: To connect your existing business intelligence tools. And query the data.

Question-145: What is Altus Share Data Experience (SDX) service?

Answer: If you are running multiple clusters and workloads and wanted to have a consistent view of the data across all the clusters. Then Altus SDX namespace externalizes cluster metadata into a shared long running service and this service is available to multiple clusters and workloads running in the public Cloud.

Question-146: Using Altus can we create cluster across multiple AWS account or Azure subscriptions?

Answer: Yes, Using the Altus Environment which identifies the resources in your AWS account or Azure subscriptions which needs to be used for Cluster or Jobs. Even using Altus Environment, you can create Clusters in multiple AWS accounts or Azure subscription even from the single Altus account.

Question-147: Can I create Cluster using Altus which support more than one Compute engine?

Answer: Yes, you can have one or more from the following compute engine

- Hive
- Spark
- Hive on Spark
- MapReduce2

Question-148: What is the Job Queue?

Answer: In the Altus Data Engineering service each cluster has a Job queue to manage the jobs that run on the cluster and supports a workflow with a single Pipeline.

Question-149: You have huge volume of structured data stored in S3, and wanted to create a Data Warehouse solution on that using CDH cluster, how can you do?

Answer: You will be using Altus Data warehouse service for such requirement using that you will be provisioning cluster in your AWS account, so all your business users would have permission. Now configure a cluster with the Impala SQL engine to enable you to interactively access your data stored in your Cloud object storage for analysis and reporting. Exactly the same can be done using Azure subscriptions.

Question-150: What is the Altus SDX Namespace?

Answer: As CDH cluster access the data stored in the public Cloud, all the metadata for this stored data is also stored in a Database. Then Altus SDX namespace points to that database and provide a common and consistent view of the data to the clusters. This SDX namespace service would be shared across multiple Altus cluster which wanted to access the same data and provide the consistent to all the cluster for these data. Actual data would be stored either in AWS S3 or Azure ADLS.

Question-151: Can I use Cloudera Altus with the Hybrid Cloud?

Answer: Yes, you can. Cloudera Altus makes it easier to extend your on-premise analytics to the Cloud.

Question-152: How could I manage the infrastructure myself for the cluster created using Cloudera Altus?

Answer: If you want to take full control of the infrastructure and manage the cluster yourself use the Altus Director.

Question-153: What are the types of metadata shared by Altus SDX service?

Answer: SDX for the Cloudera Altus makes it possible to persist and share data context such as table definition, data security, and data governance policies across long running and transient cloud workloads.

Question-154: What are all the benefits of the Cloudera Altus?

Answer: Using the Cloudera Altus you can run your Data warehouse solution, Machine Learning Jobs etc. With the following major benefits.

- No need of dedicated cluster.
- No need of dedicated Cloud resources for transient and long running data pipelines.
- Use of public cloud as of now Azure and AWS.
- Data engineers can create self-service for building and using data pipelines.

Question-155: What is the use of Cloudera Altus Director?

Answer: Cloudera Altus Director is used to provision Cloud environment for Data Engineering, Data Warehouse, Operational Database in the cloud.

Cloudera Certification & Training Material Available on HadoopExam.com are below. Please keep visiting website for that latest products



95 Q & A
[Click Here](#)



73 Q & A
[Click Here](#)



90+ Solved Scenarios
[Click Here](#)



79 Q & A
[Click Here](#)

Cloudera Hadoop
& Spark Developer
CCA175

Cloudera Hadoop
BigData Analytics
CCA159

Cloudera Hadoop
Administrator
Certification
CCA131

Cloudera Data
Engineer
CCP:DE575



[Click Here](#)



57 Solved Scenarios
[Click Here](#)

Package Deal

Hortonworks
HDPCA : Hadoop
Admin Certification