Top 100 Hadoop Interview Questions and Answers

1. What is Apache Hadoop?

Hadoop is an open source software framework for distributed storage and distributed processing of large data sets. Open source means it is freely available and even we can change its source code as per our requirements. Apache Hadoop makes it possible to run applications on the system with thousands of commodity hardware nodes. It's distributed file system has the provision of rapid data transfer rates among nodes. It also allows the system to continue operating in case of node failure.

2. Main Components of Hadoop?

Storage layer – HDFS
Batch processing engine – MapReduce
Resource Management Layer – YARN

HDFS - HDFS (Hadoop Distributed File System) is the storage unit of Hadoop. It is responsible for storing different kinds of data as blocks in a distributed environment. It follows master and slave topology.
Components of HDFS are NameNode and DataNode

MapReduce - For processing large data sets in parallel across a hadoop cluster, Hadoop MapReduce framework is used. Data analysis uses a two-step map and reduce process.

YARN - YARN (Yet Another Resource Negotiator) is the processing framework in Hadoop, which manages resources and provides an execution environment to the processes.
Main Components of YARN are Node Manager and Resource Manager

3. **Why do we need Hadoop?**

Storage – Since data is very large, so storing such huge amount of data is very difficult.
Security – Since the data is huge in size, keeping it secure is another challenge.
Analytics – In Big Data, most of the time we are unaware of the kind of data we are dealing with. So analyzing that data is even more difficult.
Data Quality – In the case of Big Data, data is very messy, inconsistent and incomplete.
Discovery – Using a powerful algorithm to find patterns and insights are very difficult.

4. What are the four characteristics of Big Data?

Volume: The volume represents the amount of data which is growing at an exponential rate i.e. in Petabytes and Exabytes.

Velocity: Velocity refers to the rate at which data is growing, which is very fast. Today, yesterday's data are considered as old data. Nowadays, social media is a major contributor in the velocity of growing data.

Variety: Variety refers to the heterogeneity of data types. In another word, the data which are gathered has a variety of formats like videos, audios, csv, etc. So, these various formats represent the variety of data.

Value: It is all well and good to have access to big data but unless we can turn it into a value it is useless.

5. What are the modes in which Hadoop run?

Local (Standalone) Mode – Hadoop by default run in a single-node, non-distributed mode, as a single Java process.

Pseudo-Distributed Mode – Just like the Standalone mode, Hadoop also runs on a single-node in a Pseudo-distributed mode.

Fully-Distributed Mode – In this mode, all daemons execute in separate nodes forming a multi-node cluster. Thus, it allows separate nodes for Master and Slave.

6. Explain about the indexing process in HDFS.

Indexing process in HDFS depends on the block size. HDFS stores the last part of the data that further points to the address where the next part of data chunk is stored.

7. What happens to a NameNode that has no data?

There does not exist any NameNode without data. If it is a NameNode then it should have some sort of data in it.

8. What is Hadoop streaming?

Hadoop distribution has a generic application programming interface for writing Map and Reduce jobs in any desired programming language like Python, Perl, Ruby, etc. This is referred to as Hadoop Streaming. Users can create and run jobs with any kind of shell scripts or executable as the Mapper or Reducers.

9. What is a block and block scanner in HDFS?

Block - The minimum amount of data that can be read or written is generally referred to as a "block" in HDFS. The default size of a block in HDFS is 64MB.

Block Scanner - Block Scanner tracks the list of blocks present on a DataNode and verifies them to find any kind of checksum errors. Block Scanners use a throttling mechanism to reserve disk bandwidth on the datanode.

10. What is a checkpoint?

Checkpoint Node keeps track of the latest checkpoint in a directory that has same structure as that of NameNode's directory. Checkpoint node creates checkpoints for the namespace at regular intervals by downloading the edits and fsimage file from the NameNode and merging it locally. The new image is then again updated back to the active NameNode.

11. What is commodity hardware?

Commodity Hardware refers to inexpensive systems that do not have high availability or high quality. Commodity Hardware consists of RAM because there are specific services that need to be executed on RAM. Hadoop can be run on any commodity hardware and does not require any super computer s or high end hardware configuration to execute jobs.

12. Explain what is heartbeat in HDFS?

Heartbeat is referred to a signal used between a data node and Name node, and between task tracker and job tracker, if the Name node or job tracker does not respond to the signal, then it is considered there is some issues with data node or task tracker.

13. What happens when a datanode fails ?

When a datanode fails
Jobtracker and namenode detect the failure
On the failed node all tasks are re-scheduled
Namenode replicates the users data to another node

14. Explain what happens in textinformat ?

In textinputformat, each line in the text file is a record. Value is the content of the line while Key is the byte offset of the line. For instance, Key: longWritable, Value: text

15. Explain what is sqoop in Hadoop ?

To transfer the data between Relational database management (RDBMS) and Hadoop HDFS a tool is used known as Sqoop. Using Sqoop data can be transferred from RDMS like MySQL or Oracle into HDFS as well as exporting data from HDFS file to RDBMS.

16. Mention what are the data components used by Hadoop?

Data components used by Hadoop are

Pig
Hive

17. **What is rack awareness?**
    Rack awareness is the way in which the namenode determines on how to place blocks based on the rack definitions.

18. **Explain how do 'map' and 'reduce' works.**
    Namenode takes the input and divide it into parts and assign them to data nodes. These datanodes process the tasks assigned to them and make a key-value pair and returns the intermediate output to the Reducer. The reducer collects this key value pairs of all the datanodes and combines them and generates the final output.

19. **What is a Combiner?**
    The Combiner is a 'mini-reduce' process which operates only on data generated by a mapper. The Combiner will receive as input all data emitted by the Mapper instances on a given node. The output from the Combiner is then sent to the Reducers, instead of the output from the Mappers.

20. **Consider case scenario: In M/R system, - HDFS block size is 64 MB**
    - Input format is FileInputFormat
    – We have 3 files of size 64K, 65Mb and 127Mb
    How many input splits will be made by Hadoop framework?

Hadoop will make 5 splits as follows −
- 1 split for 64K files
- 2 splits for 65MB files
- 2 splits for 127MB files

21. **Suppose Hadoop spawned 100 tasks for a job and one of the task failed. What will Hadoop do?**
    It will restart the task again on some other TaskTracker and only if the task fails more than four (the default setting and can be changed) times will it kill the job.

22. **What are Problems with small files and HDFS?**
    HDFS is not good at handling large number of small files. Because every file, directory and block in HDFS is represented as an object in the namenode's memory, each of which occupies approx 150 bytes So 10 million files, each using a block, would use about 3 gigabytes of memory. when we go for a billion files the memory requirement in namenode cannot be met.

23. What does 'jps' command do?

It gives the status of the deamons which run Hadoop cluster. It gives the output mentioning the status of namenode, datanode , secondary namenode, Jobtracker and Task tracker.

24. How to restart Namenode?

Step-1. Click on stop-all.sh and then click on start-all.sh OR

Step-2. Write sudo hdfs (press enter), su-hdfs (press enter), /etc/init.d/ha (press enter) and then /etc/init.d/hadoop-0.20-namenode start (press enter).

25. What does /etc /init.d do?

/etc /init.d specifies where daemons (services) are placed or to see the status of these daemons. It is very LINUX specific, and nothing to do with Hadoop.

26. Mention what is the use of Context Object?

The Context Object enables the mapper to interact with the rest of the Hadoop system. It includes configuration data for the job, as well as interfaces which allow it to emit output.

27. Mention what is the number of default partitioner in Hadoop?

In Hadoop, the default partitioner is a "Hash" Partitioner.

28. Explain what is the purpose of RecordReader in Hadoop?

In Hadoop, the RecordReader loads the data from its source and converts it into (key, value) pairs suitable for reading by the Mapper.

29. Mention what is the best way to copy files between HDFS clusters?

The best way to copy files between HDFS clusters is by using multiple nodes and the distcp command, so the workload is shared.

30. What is "speculative execution" in Hadoop?

If a node appears to be executing a task slower, the master node can redundantly execute another instance of the same task on another node. Then, the task which finishes first will be accepted and the other one is killed. This process is called "speculative execution".

31. Explain what is difference between an Input Split and HDFS Block?

Logical division of data is known as Split while physical division of data is known as HDFS Block.

32. How can native libraries be included in YARN jobs?

There are two ways to include native libraries in YARN jobs-

1) By setting the -Djava.library.path on the command line but in this case there are chances that the native libraries might not be loaded correctly and there is possibility of errors.

2) The better option to include native libraries is to the set the LD_LIBRARY_PATH in the .bashrc file.

33. What is Apache HBase?

HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java. HBase runs on top of HDFS (Hadoop Distributed File System) and provides BigTable (Google) like capabilities to Hadoop. It is designed to provide a fault tolerant way of storing large collection of sparse data sets. HBase achieves high throughput and low latency by providing faster Read/Write Access on huge data sets.

34. What is "SerDe" in "Hive"?

Apache Hive is a data warehouse system built on top of Hadoop and is used for analyzing structured and semi-structured data developed by Facebook. Hive abstracts the complexity of Hadoop MapReduce.

The "SerDe" interface allows you to instruct "Hive" about how a record should be processed. A "SerDe" is a combination of a "Serializer" and a "Deserializer". "Hive" uses "SerDe" (and "FileFormat") to read and write the table's row.

35. Explain "WAL" in HBase?

Write Ahead Log (WAL) is a file attached to every Region Server inside the distributed environment. The WAL stores the new data that hasn't been persisted or committed to the permanent storage. It is used in case of failure to recover the data sets.

36. What is Apache Spark?

The answer to this question is, Apache Spark is a framework for real time data analytics in a distributed computing environment. It executes in-memory computations to increase the speed of data processing.

It is 100x faster than MapReduce for large scale data processing by exploiting in-memory computations and other optimizations.

37. What is a UDF?

If some functions are unavailable in built-in operators, we can programmatically create User Defined Functions (UDF) to bring those functionalities using other languages like Java, Python, Ruby, etc. and embed it in Script file.

38. Explain about the SMB Join in Hive.

In SMB join in Hive, each mapper reads a bucket from the first table and the corresponding bucket from the second table and then a merge sort join is performed. Sort Merge Bucket (SMB) join in hive is mainly used as there is no limit on file or partition or table join. SMB join can best be used when the tables are large. In SMB join the columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

39. How can you connect an application, if you run Hive as a server?

When running Hive as a server, the application can be connected in one of the 3 ways-
ODBC Driver-This supports the ODBC protocol
JDBC Driver- This supports the JDBC protocol
Thrift Client- This client can be used to make calls to all hive commands using different programming language like PHP, Python, Java, C++ and Ruby.

40. Is YARN a replacement of Hadoop MapReduce?

YARN is not a replacement of Hadoop but it is a more powerful and efficient technology that supports MapReduce and is also referred to as Hadoop 2.0 or MapReduce 2.

41. What is a Record Reader?

A RecordReader uses the data within the boundaries created by the input split to generate key/value pairs. Each of the generated Key/value pair will be sent one by one to their mapper.

42. What is a sequence file in Hadoop?

Sequence file is used to store binary key/value pairs. Sequence files support splitting even when the data inside the file is compressed which is not possible with a regular compressed file. You can either choose to perform a record level compression in which the value in the key/value pair will be compressed. Or you can also choose to choose at the block level where multiple records will be compressed together.

43. How do you overwrite replication factor?

There are few ways to do this. Look at the below illustration.

Illustration
hadoop fs -setrep -w 5 -R hadoop-test
hadoop fs -Ddfs.replication=5 -cp hadoop-test/test.csv hadoop-test/test_with_rep5.csv

44. How do you do a file system check in HDFS?

FSCK command is used to do a file system check in HDFS. It is a very useful command to check the health of the file, block names and block locations.

Illustration
hdfs fsck /dir/hadoop-test -files -blocks -locations

45. Is Namenode also a commodity?

No. Namenode can never be a commodity hardware because the entire HDFS rely on it. It is the single point of failure in HDFS. Namenode has to be a high-availability machine.

46. What is the difference between an InputSplit and a Block?

Block is a physical division of data and does not take in to account the logical boundary of records. Meaning you could have a record that started in one block and ends in another block. Where as InputSplit considers the logical boundaries of records as well.

47. What is the difference between SORT BY and ORDER BY in Hive?

ORDER BY performs a total ordering of the query result set. This means that all the data is passed through a single reducer, which may take an unacceptably long time to execute for larger data sets.

SORT BY orders the data only within each reducer, thereby performing a local ordering, where each reducer's output will be sorted. You will not achieve a total ordering on the dataset. Better performance is traded for total ordering.

48. In which directory Hadoop is installed?

Cloudera and Apache has the same directory structure. Hadoop is installed in cd/usr/lib/hadoop/

49. What are the port numbers of Namenode, job tracker and task tracker?

The port number for Namenode is '50070', for job tracker is '50030' and for task tracker is '50060'.

50. What are the Hadoop configuration files at present?

There are 3 configuration files in Hadoop:

1.core-site.xml
2.hdfs-site.xml
3.mapred-site.xml
These files are located in thehadoop/conf/subdirectory.

51. What is Cloudera and why it is used?

Cloudera is the distribution of Hadoop. It is a user created on VM by default. Cloudera belongs to Apache and is used for data processing.

52. How can we check whether Namenode is working or not?

To check whether Namenode is working or not, use the command /etc/init.d/hadoop-namenode status.

53. Which files are used by the startup and shutdown commands?

Slaves and Masters are used by the startup and the shutdown commands.

54. Can we create a Hadoop cluster from scratch?

Yes we can do that also once we are familiar with the Hadoop environment.

55. How can you transfer data from Hive to HDFS?

By writing the query:
hive> insert overwrite directory '/' select * from emp;

You can write your query for the data you want to import from Hive to HDFS. The output you receive will be stored in part files in the specified HDFS path.

56. What is Job Tracker role in Hadoop?

Job Tracker's primary function is resource management (managing the task trackers), tracking resource availability and task life cycle management (tracking the taks progress and fault tolerance).

- It is a process that runs on a separate node, not on a DataNode often.
- Job Tracker communicates with the NameNode to identify data location.
- Finds the best Task Tracker Nodes to execute tasks on given nodes.
- Monitors individual Task Trackers and submits the overall job back to the client.
- It tracks the execution of MapReduce workloads local to the slave node.

57. What are the core methods of a Reducer?

The three core methods of a Reducer are:

setup(): this method is used for configuring various parameters like input data size, distributed cache.
public void setup (context)

reduce(): heart of the reducer always called once per key with the associated reduced task
public void reduce(Key, Value, context)

cleanup(): this method is called to clean temporary files, only once at the end of the task
public void cleanup (context)

58. Compare Hadoop & Spark

Criteria	**Hadoop**	**Spark**
Dedicated storage	HDFS	None
Speed of processing	Average	Excellent
Libraries	Separate tools available	Spark Core, SQL, Streaming, MLlib, GraphX

59. Can i access Hive Without Hadoop ?

Yes,We can access Hive without hadoop with the help of other data storage systems like Amazon S3, GPFS (IBM) and MapR file system .

60. What is Apache Spark?

Spark is a fast, easy-to-use and flexible data processing framework. It has an advanced execution engine supporting cyclic data flow and in-memory computing. Spark can run on Hadoop, standalone or in the cloud and is capable of accessing diverse data sources including HDFS, HBase, Cassandra and others.

61. How Spark uses Hadoop?

Spark has its own cluster management computation and mainly uses Hadoop for storage.

62. What is Spark SQL?

SQL Spark, better known as Shark is a novel module introduced in Spark to work with structured data and perform structured data processing. Through this module, Spark executes relational SQL queries on the data. The core of the component supports an altogether different RDD called SchemaRDD, composed of rows objects and schema objects defining data type of each column in the row. It is similar to a table in relational database.

63. What are the additional benefits YARN brings in to Hadoop?

Effective utilization of the resources as multiple applications can be run in YARN all sharing a common resource.YARN is backward compatible so all the existing MapReduce jobs.Using YARN, one can even run applications that are not based on the MaReduce model

64. Compare Sqoop and Flume

Criteria	**Sqoop**	**Flume**
Application	Importing data from RDBMS	Moving bulk streaming data into HDFS
Architecture	Connector – connecting to respective data	Agent – fetching of the right data
Loading of data	Event driven	Not event driven

65. What is Sqoop metastore?

Sqoop metastore is a shared metadata repository for remote users to define and execute saved jobs created using sqoop job defined in the metastore. The sqoop –site.xml should be configured to connect to the metastore.

66. Which are the elements of Kafka?

The most important elements of Kafka:

Topic – It is the bunch of similar kind of messages

Producer – using this one can issue communications to the topic

Consumer – it endures to a variety of topics and takes data from brokers.

Brokers – this is the place where the issued messages are stored

67. What is Kafka?

Wikipedia defines Kafka as “an open-source message broker project developed by the Apache Software Foundation written in Scala, where the design is heavily influenced by transaction logs”. It is essentially a distributed publish-subscribe messaging system.

68. What is the role of the ZooKeeper?

Kafka uses Zookeeper to store offsets of messages consumed for a specific topic and partition by a specific Consumer Group.

69. What are the key benefits of using Storm for Real Time Processing?

Easy to operate : Operating storm is quiet easy.

Real fast : It can process 100 messages per second per node.

Fault Tolerant : It detects the fault automatically and re-starts the functional attributes.

Reliable : It guarantees that each unit of data will be executed at least once or exactly once.

Scalable : It runs across a cluster of machine

70. List out different stream grouping in Apache storm?

- Shuffle grouping
- Fields grouping
- Global grouping
- All grouping
- None grouping
- Direct grouping
- Local grouping

71. Which operating system(s) are supported for production Hadoop deployment?

The main supported operating system is Linux. However, with some additional software Hadoop can be deployed on Windows.

72. What is the best practice to deploy the secondary namenode

Deploy secondary namenode on a separate standalone machine. The secondary namenode needs to be deployed on a separate machine. It will not interfere with primary namenode operations in this way. The secondary namenode must have the same memory requirements as the main namenode.

73. What are the side effects of not running a secondary name node?

The cluster performance will degrade over time since edit log will grow bigger and bigger. If the secondary namenode is not running at all, the edit log will grow significantly and it will slow the system down. Also, the system will go into safemode for an extended time since the namenode needs to combine the edit log and the current filesystem checkpoint image.

74. What daemons run on Master nodes?

NameNode, Secondary NameNode and JobTracker

Hadoop is comprised of five separate daemons and each of these daemon run in its own JVM. NameNode, Secondary NameNode and JobTracker run on Master nodes. DataNode and TaskTracker run on each Slave nodes.

75. Explain about the BloomMapFile.

BloomMapFile is a class, that extends the MapFile class. It is used in HBase table format to provide quick membership test for the keys using dynamic bloom filters.

76. What is the usage of foreach operation in Pig scripts?

FOREACH operation in Apache Pig is used to apply transformation to each element in the data bag, so that respective action is performed to generate new data items.

Syntax- FOREACH data_bagname GENERATE exp1, exp2

77. Explain about the different complex data types in Pig.

Apache Pig supports 3 complex data types-

Maps- These are key, value stores joined together using #.

Tuples- Just similar to the row in a table, where different items are separated by a comma. Tuples can have multiple attributes.

Bags- Unordered collection of tuples. Bag allows multiple duplicate tuples.

78. Differentiate between PigLatin and HiveQL

- It is necessary to specify the schema in HiveQL, whereas it is optional in PigLatin.
- HiveQL is a declarative language, whereas PigLatin is procedural.
- HiveQL follows a flat relational data model, whereas PigLatin has nested relational data model.

79. Whether pig latin language is case-sensitive or not?

Answer: pig latin is sometimes not a case sensitive.let us see example,Load is equivalent to load.

A=load ‘b’ is not equivalent to a=load ‘b’

UDF are also case sensitive,count is not equivalent to COUNT.

80. What are the use cases of Apache Pig?

Apache Pig is used for analyzing and performing tasks involving ad-hoc processing. Apache Pig is used for:

Research on large raw data sets like data processing for search platforms. For example, Yahoo uses Apache Pig to analyse data gathered from Yahoo search engines and Yahoo News Feeds.

Processing huge data sets like Web logs, streaming online data, etc.

In customer behavior prediction models like e-commerce websites.

81. What does Apache Mahout do?

Mahout supports four main data science use cases:

Collaborative filtering – mines user behavior and makes product recommendations (e.g. Amazon recommendations)

Clustering – takes items in a particular class (such as web pages or newspaper articles) and organizes them into naturally occurring groups, such that items belonging to the same group are similar to each other

Classification – learns from existing categorizations and then assigns unclassified items to the best category

Frequent item-set mining – analyzes items in a group (e.g. items in a shopping cart or terms in a query session) and then identifies which items typically appear together

82. Mention some machine learning algorithms exposed by Mahout?

Below is a current list of machine learning algorithms exposed by Mahout.

Collaborative Filtering

- Item-based Collaborative Filtering
- Matrix Factorization with Alternating Least Squares
- Matrix Factorization with Alternating Least Squares on Implicit Feedback

Classification

- Naive Bayes
- Complementary Naive Bayes
- Random Forest

Clustering

- Canopy Clustering
- k-Means Clustering
- Fuzzy k-Means
- Streaming k-Means
- Spectral Clustering

83. What is Apache Flume?

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data source. Review this Flume use case to learn how Mozilla collects and Analyse the Logs using Flume and Hive.

Flume is a framework for populating Hadoop with data. Agents are populated throughout ones IT infrastructure – inside web servers, application servers and mobile devices, for example – to collect data and integrate it into Hadoop.

84. Explain about the different channel types in Flume. Which channel type is faster?

The 3 different built in channel types available in Flume are-

MEMORY Channel – Events are read from the source into memory and passed to the sink.

JDBC Channel – JDBC Channel stores the events in an embedded Derby database.

FILE Channel –File Channel writes the contents to a file on the file system after reading the event from a source. The file is deleted only after the contents are successfully delivered to the sink.

MEMORY Channel is the fastest channel among the three however has the risk of data loss. The channel that you choose completely depends on the nature of the big data application and the value of each event.

85. Why we are using Flume?

Most often Hadoop developer use this too to get data from social media sites. Its developed by Cloudera for aggregating and moving very large amount if data. The primary use is to gather log files from different sources and asynchronously persist in the hadoop cluster.

86. Which Scala library is used for functional programming?

Scalaz library has purely functional data structures that complement the standard Scala library. It has pre-defined set of foundational type classes like Monad, Functor, etc.

87. What do you understand by "Unit" and "()" in Scala?

Unit is a subtype of scala.anyval and is nothing but Scala equivalent of Java void that provides the Scala with an abstraction of the java platform. Empty tuple i.e. () in Scala is a term that represents unit value.

88. What do you understand by a closure in Scala?

Closure is a function in Scala where the return value of the function depends on the value of one or more variables that have been declared outside the function.

89. List some use cases where classification machine learning algorithms can be used.

- Natural language processing (Best example for this is Spoken Language Understanding)
- Market Segmentation
- Text Categorization (Spam Filtering)
- Bioinformatics (Classifying proteins according to their function)
- Fraud Detection
- Face detection

90. Mention what is data cleansing?

Data cleaning also referred as data cleansing, deals with identifying and removing errors and inconsistencies from data in order to enhance the quality of data.

91. List of some best tools that can be useful for data-analysis?

- Tableau
- RapidMiner
- OpenRefine
- KNIME
- Google Search Operators
- Solver
- NodeXL
- io
- Wolfram Alpha's
- Google Fusion tables

92. List out some common problems faced by data analyst?

Some of the common problems faced by data analyst are

- Common misspelling
- Duplicate entries
- Missing values
- Illegal values
- Varying value representations
- Identifying overlapping data

93. Explain what are the tools used in Big Data?

- Hadoop
- Hive
- Pig
- Flume
- Mahout
- Sqoop

94. Which language is more suitable for text analytics? R or Python?

Since Python consists of a rich library called Pandas which allows the analysts to use high-level data analysis tools as well as data structures, while R lacks this feature. Hence Python will more suitable for text analytics.

95. What is logistic regression?

It is a statistical technique or a model in order to analyze a dataset and predict the binary outcome. The outcome has to be a binary outcome that is either zero or one or a yes or no.

96. Can you list few commonly used Hive services?

- Command Line Interface (cli)
- Hive Web Interface (hwi)
- HiveServer (hiveserver)
- Printing the contents of an RC file using the tool rcfilecat.
- Jar
- Metastore

97. What is indexing and why do we need it?

One of the Hive query optimization methods is Hive index. Hive index is used to speed up the access of a column or set of columns in a Hive database because with the use of index the database system does not need to read all rows in the table to find the data that one has selected.

98. What are the components used in Hive query processor?

The components of a Hive query processor include

- Logical Plan of Generation.
- Physical Plan of Generation.
- Execution Engine.
- Operators.
- UDF's and UDAF's.
- Optimizer.
- Parser.
- Semantic Analyzer.
- Type Checking

99. If you run a select * query in Hive, Why does it not run MapReduce?

The hive.fetch.task.conversion property of Hive lowers the latency of mapreduce overhead and in effect when executing queries like SELECT, FILTER, LIMIT, etc., it skips mapreduce function.

100. What is the use of explode in Hive?

Explode in Hive is used to convert complex data types into desired table formats. explode UDTF basically emits all the elements in an array into multiple rows.