27/5/2022

Takhti
Date
Page

Name: Vicky Kumar
Roll no: 18ERECS080
Subject: Big Data Analysis.
(8th Semester) - (2nd Mid-Term)
Sec-A

Ans① The OODA loop (Observe, Orient, Decide, Act) is a four step approach to decision-making that focuses on filtering available information, putting it in context and quickly making the most approach decision while also understanding that chances are can be made as more data become available.

Ans② The tableau workspace consists of menus, a toolbar, the Data pane, cards and shelves, and one or more sheets. Sheets can be worksheets, dashboards or stories.

Ans③ Visual options in tableau.

(i) Bar graph	(v) Density Maps
(ii) Line chart	(vi) Scatter plot
(iii) Pie chart	(vii) Gantt chart
(iv) Maps	(viii) Bubble chart.

Takhti
Date
Page

Ans(4) Mapreduce is a programming paradigm that enables massive scalability accross hundreds or thousands of servers in a Hadoop clusters.

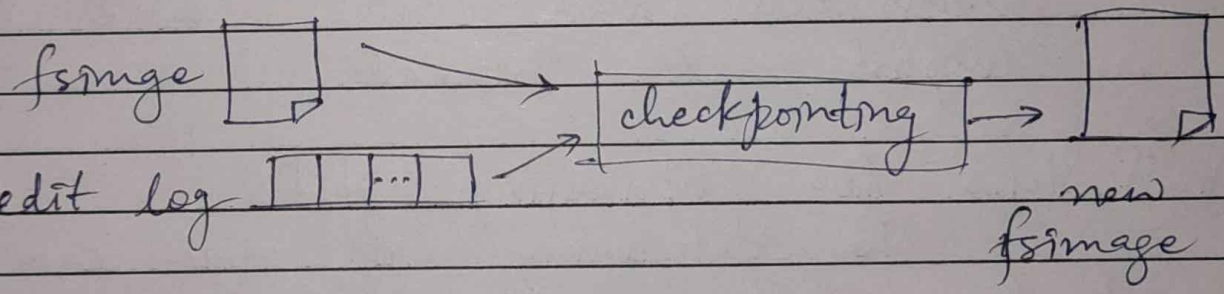
Ans(5) A key-value database is a type of non-relational database that uses a simple key value method to store data.
A key-value database stores data as a collection of key value pairs in which a key serves as a unique identifier.

(Sec-B)

Ans(2) A typical edit ranges from 10s to 100s of bytes, but over time enough edits can accumulate to become unwidely.
In extreme cases, it can fill up all the available disk capacity on a node, but more subtly, a large edit log can substantially delay Namenode startup.

Takhti
Date
Page

Checkpoints is a process that takes an fsimage and edit log and compacts them into a new fsimage.
This way, instead of replying a potentially unbounded edit log, the NameNode can load the final in-memory state directly from the fsimage. This is far more efficient operation and reduce NameNode startup time.



Ans: (3) XML files Configuration :-

(i) Configuration Options for XML files.
(ii) Basic Configuration
(iii) Advanced Configuration
(iv) Content Segmentation.

The XML files similar to CSV and XLSX files, requires

Takhti
Date
Page

additional configuration after uploading to the project so the system could import the content of these files.

↳ To configure XML files,

i) Open your project and go to Content → Files.
ii) Click Configure next to the files open the configuration window.
iii) Select which content should be translated and click Save & import to proceed.

Ans (4) Steps for upgrading HDFS.

To determine if you can finalize the upgrade, run important workloads and ensure that they are successful.

Make sure you have enough free disk space, keeping in mind that,

i) Deleting files does not free up disk files.

(ii) Using the balancer causes all moved or replicas to be duplicated.

If you have Enabled high availability for HDFS, and you have performed a rolling upgrade.

(i) Go to the HDFS services.
(ii) Select Action → Finalize rolling upgrade and click Finalize Rolling Upgrade to confirm.

If you have not performed a rolling upgrade:

(i) Go to the HDFS services.
(ii) Click the Instances tab.
(iii) Click the link for the Instant NameNode instance.

Ans ⑥ Data Visualization Techniques :-

↳ Box plots
↳ Histograms
↳ Heat maps
↳ Charts
↳ Tree maps.

Takhti
Date
Page

* Lists of Methods to Visualize Data :-

(i) Column chart
(ii) Bar chart
(iii) Stacked Bar graph
(iv) Stacked Columns chart.
(v) Area chart
(vi) Dual Axis chart
(vii) Line graph
(viii) Mekko chart
(ix) Pie chart
(x) Waterfall Chart
(xi) Bubble chart
(xii) Scatter Plot chart
(xiii) Bullet Graph
(xiv) Funnel chart
(xv) Heat map.

(Sec-C)

Ans 1. Hive DML (Data Manipulation language) commands are used to insert, update, retrieve and delete data from the Hive table once the table and database schema has been defined using Hive DDL commands.

Takhti
Date
Page

The various Hive DML commands are:

(i) LOAD
(ii) SELECT
(iii) INSERT
(iv) DELETE
(v) UPDATE
(vi) EXPORT
(vii) IMPORT

(a) The LOAD commands in Hive is used to move data files into the locations corresponding to Hive table.

If local keyword is specified, then the LOAD command will look for the path in local filesystem.

(b) SELECT commands in Hive is similar to the SELECT statement in SQL used for retrieve data from the database.

(c) INSERT INTO statement appends the data into existing

da

Takhti
Date
Page

data in the table or partition.

(d) The DELETE statement in Hive deletes the table data.
If the WHERE clause is specified, then it deletes the rows that specify the condition in where clause.

(e) UPDATE :- It can perform on the hive table that support ACID.

(f) The EXPORT statement in Hive exports the table to specified location in the HDFS.

(g) The Hive IMPORT command imports the data from a specified location to a new table or already existing tables.

Ans.(2) Five characteristics of Big Data :-

Big Data is a collection of data from many different sources and is often describe by five characteristics :-

Takhti
Date
Page

* Volume, Value, Variety, Velocity and Veracity.

(a) Volume :- The size and amount of big-data that companies manage and analyse.

(b) Value :- The most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and qualificable business benefits.

(c) Variety :- The diversity and range of different data types, including unstructured data, semi-structured data and raw data.

(d) Velocity :- The speed of companies receive, store and manage data - e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time.

Takhti
Date
Page

e) Veracity :- The "truth" or accuracy of data and infrastructure assests, which often determines executive-level confidence.

The additional characteristics of variability can also be considered.

↳ Variability :- The changing nature of data companies seek to capture, manage and analyse - eg:- in sentiment or text analysis, changes in the meaning of key words or phrases.
