**Date : 28/05/2022**

# Assignment – 02

# Big Data Analytics

| |
|---|
| Y. S. V. SUMANTH |
| 19131A05R6 |
| CSE 4 |

---

**Q1) Distinguish between Traditional Data Analysis and Big Data Analysis.**

**A) Traditional Data Analysis :**

- Traditional data analysis means to use proper statistical methods to analyse massive first-hand data and second-hand data, to concentrate, extract, and refine useful data hidden in a batch of chaotic data, and to identify the inherent law of the subject matter, so as to develop functions of data to the greatest extent and maximize the value of data.
- Traditional data analysis occurs incrementally i.e. an event occurs, data is generated, and the analysis of this data takes place after the event.
- Traditional data analysis can help businesses understand the impacts of given strategies or changes on a limited range of metrics over a specific period.
- Traditional data analysis can be more narrow and too restricted to deliver the meaningful insights big data can provide.
- Traditional Data Analysis can be easier to secure, which may make it preferable for highly sensitive, personal, or confidential data sets. Because traditional data is smaller, it doesn't require distributed architecture and is less likely to require third-party storage.
- Traditional Data Analysis can be processed using conventional data processing software and a normal system configuration. Processing big data generally requires a higher-configuration setup, which can increase resource usage and costs unnecessarily when traditional data methods will suffice.
- Traditional Data Analysis is easier to manipulate and interpret. Because traditional data is simpler and relational in nature, it can be processed using normal functions and may even be accessible to nonexperts.

**Big Data Analysis :**

- Big data analysis can be deemed as the analysis of a special kind of data. Therefore, many traditional data analysis methods may still be utilized for big data analysis.
- Big data analysis can occur in real time. Because big data generates on a second-by-second basis, analysis can occur as data is being collected.
- Big data analysis offers businesses a more dynamic, holistic understanding of their needs and strategies.
- Big data analysis can provide a deeper analysis of market trends and consumer behaviour. Traditional data analysis can be more narrow and too restricted to deliver the meaningful insights big data can provide.
- Big data analysis provides insights faster. Organizations can learn from big data in real time. In the context of big data analytics, this can provide a competitive edge.
- Big data analysis is more efficient. The increasingly digital nature of our society means people and businesses are generating vast quantities of data every day—and even every minute. Big data allows us to harness this data and interpret it in a meaningful way.

- Big data analysis requires advanced preparation. To leverage the benefits, organizations need to prepare for big data through new security protocols, configuration steps, increases in available processing power.

**Q2) What are the big data analytical methods ? Explain in detail.**

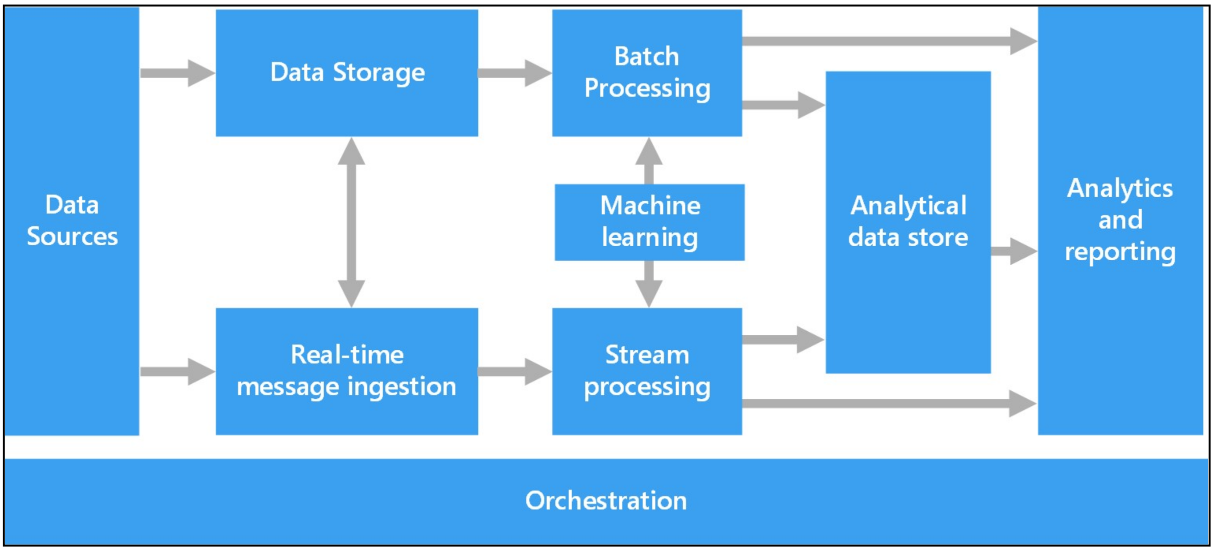
**A)** In the dawn of the big data era, people are concerned with how to rapidly extract key information from massive data so as to bring values for enterprises and individuals. At present, the main processing methods of big data are shown as follows.

1) **Bloom Filter:**
   - Bloom Filter is actually a bit array and a series of Hash functions.
   - The principle of Bloom Filter is to store Hash values of data other than data itself by utilizing a bit array, which in essence a bitmap index that uses Hash functions to conduct lossy compression storage of data.
   - It has such advantages as high space efficiency and high query speed, but also with some disadvantages like having a certain misrecognition rate and deletion difficulty.
   - Bloom Filter applies to big data applications that allow a certain misrecognition rate.
2) **Hashing:**
   - It is a method that essentially transforms data into shorter fixed-length numerical values or index values.
   - Hashing has such advantages as rapid reading, writing, and high query speed, but a sound Hash function is hard to be found.
3) **Index:**
   - Index is always an effective method to reduce the expense of disc reading and writing, and improve insertion, deletion, modification, and query speeds in both traditional relational databases that manage structured data, and technologies that manage semi-structured and unstructured data.
   - However, index has a disadvantage that it has the additional cost for storing index files and the index files should be maintained dynamically according to data updates.
4) **Triel:**
   - It is also called Trie tree, a variant of Hash Tree.
   - It is mainly applied to rapid retrieval and word frequency statistics.
   - The main idea of Triel is to utilize common prefixes of character strings to reduce comparison on character strings to the greatest extent, so as to improve query efficiency.
5) **Parallel Computing:**
   - Compared to traditional serial computing, parallel computing refers to utilizing several computing resources to complete a computation task.
   - Its basic idea is to decompose a problem and assign them to several independent processes to be independently completed, so as to achieve coprocessing.
   - Presently, some classic parallel computing models include MPI (Message Passing Interface), MapReduce, and Dryad.

**Q3) Draw the Architecture for Big Data Analysis.**

## A) **Architecture for Big Data Analysis :**

Due to the wide range of sources and variety, different structures, and the broad application fields of big data, different analytical architectures shall be considered for big data with different application requirements.



**Big Data Analysis Architecture**

Big data solutions typically involve one or more of the following types of workload :

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.

Most big data architectures include some or all of the following components :

**1) Data sources :** All big data solutions start with one or more data sources. Examples include :

- Application data stores, such as relational databases.
- Static files produced by applications, such as web server log files.
- Real-time data sources, such as IoT devices.

**2) Data storage :**

- Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats.
- This kind of store is often called a data lake. Options for implementing this storage include Azure Data Lake Store or blob containers in Azure Storage.

**3) Batch processing :**

- Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis.
- Usually these jobs involve reading source files, processing them, and writing the output to new files.

- Options include running U-SQL jobs in Azure Data Lake Analytics, using Hive, Pig, or custom Map/Reduce jobs in an HDInsight Hadoop cluster, or using Java, Scala, or Python programs in an HDInsight Spark cluster.

**4) Real-time message ingestion :**

- If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing.
- This might be a simple data store, where incoming messages are dropped into a folder for processing.
- However, many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics. Options include Azure Event Hubs, Azure IoT Hubs, and Kafka.

**5) Stream processing :**

- After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis.
- The processed stream data is then written to an output sink. Azure Stream Analytics provides a managed stream processing service based on perpetually running SQL queries that operate on unbounded streams.
- We can also use Open source Apache streaming technologies like Storm and Spark Streaming in an HDInsight cluster.

**6) Analytical data store :**

- Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools.
- The analytical data store used to serve these queries can be a Kimball-style relational data warehouse, as seen in most traditional business intelligence (BI) solutions.
- Alternatively, the data could be presented through a low-latency NoSQL technology such as HBase, or an interactive Hive database that provides a metadata abstraction over data files in the distributed data store. Azure Synapse Analytics provides a managed service for large-scale, cloud-based data warehousing.
- HDInsight supports Interactive Hive, HBase, and Spark SQL, which can also be used to serve data for analysis.

**7) Analysis and reporting** :

- The goal of most big data solutions is to provide insights into the data through analysis and reporting.
- To empower users to analyse the data, the architecture may include a data modelling layer, such as a multidimensional OLAP cube or tabular data model in Azure Analysis Services.
- It might also support self-service BI, using the modelling and visualization technologies in Microsoft Power BI or Microsoft Excel.
- Analysis and reporting can take form of interactive data exploration by data scientists or data analysts.
- For these scenarios, many Azure services support analytical notebooks, such as Jupyter, enabling these users to leverage their existing skills with Python or R.

- For large-scale data exploration, you can use Microsoft R Server, either standalone or with Spark.

**8) Orchestration :**

- Most big data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard.
- To automate these workflows, you can use an orchestration technology such Azure Data Factory or Apache Oozie and Sqoop.

## Analysis at Different Levels :

Big data analysis can also be classified into memory level analysis, Business Intelligence (BI) level analysis, and massive level analysis, which are examined in the following.

### Memory-Level :

- Memory-level analysis is for the case when the total data volume is within the maximum level of the memory of a clusters.
- The memory of current server cluster surpasses hundreds of GB while even the TB level is common.
- Therefore, an internal database technology may be used and hot data shall reside in the memory so as to improve the analytical efficiency.
- Memory-level analysis is extremely suitable for real-time analysis. MongoDB is a representative memory-level analytical architecture.
- With the development of SSD (Solid-State Drive), the capacity and performance of memory-level data analysis has been further improved and widely applied.

### BI – level :

- BI analysis is for the case when the data scale surpasses the memory level but may be imported into the BI analysis environment.
- Currently, mainstream BI products are provided with data analysis plans supporting the level over TB.

### Massive – level :

- Massive analysis for the case when the data scale has completely surpassed the capacities of BI products and traditional relational databases.
- At present, most massive analysis utilize HDFS of Hadoop to store data and use MapReduce for data analysis.
- Most massive analysis belongs to the offline analysis category.

## Analysis with Different Complexity :

- The time and space complexity of data analysis algorithms differ greatly from each other according to different kinds of data and application demands.

- For example, for applications that are amenable to parallel processing, a distributed algorithm may be designed and a parallel processing model may be used for data analysis.

**Q4) Distinguish between Real-Time Analysis and Offline Analysis ?**

**A)** Big data analysis can be classified into real-time analysis and off-line analysis according to the real-time requirement.

**Real – Time Analysis :**

**1)** Real-time analysis is mainly used in Ecommerce and finance.
**2)** Since data constantly changes, rapid data analysis is needed and analytical results shall be returned with a very short delay.
**3)** The main existing Big Data Analysis architectures of real-time analysis include
  **a)** parallel processing clusters using traditional relational databases, and
  **b)** memory-based computing platforms.
**4)** For example, Greenplum from EMC and HANA from SAP are all real-time analysis architectures.

**Offline Analysis :**

**1)** Offline analysis is usually used for applications without high requirements on response time.

**E.g. :** Machine learning, statistical analysis, and Recommendation algorithms.

**2)** Offline analysis generally conducts analysis by importing big data of logs into a special platform through data acquisition tools.
**3)** Under the big data setting, many Internet enterprises utilize the offline analysis architecture based on Hadoop in order to
  **a)** Reduce the cost of data format conversion and
  **b)** Improve the efficiency of data acquisition.
**4)** **Examples** include **1)** Facebook's Open Source tool Scribe, **2)** LinkedIn's Open Source tool Kafka, **3)** Taobao's Open source tool Time tunnel, and **4)** Chukwa of Hadoop, etc.
**5)** These tools can meet the demands of data acquisition and transmission with hundreds of MB per second.

**Q5) What are the tools required for Big Data Mining and Analysis ?**

**A) Tools for Big Data Mining and Analysis :**

Many tools for big data mining and analysis are available, including professional and amateur software, expensive commercial software, and free Open Source software.

**1) R (30.7 %) :**

- R, an open source programming language and software environment, is designed for data mining/analysis and visualization.

- While compute-intensive tasks are executed, code programmed with C, C++, and Fortran may be in under the R environment. In addition, skilled users may directly call R objects in C.
- R is a realization of the S language. S is an interpreted language developed by AT&T Bell Labs and used for data exploration, statistical analysis, and drawing plots.
- Initially, S was mainly implemented in S-PLUS, but S-PLUS is a commercial software. Compared to S, R is more popular since it is open source.
- R ranks top 1 in the KD Nuggets 2012 survey. Furthermore, in a survey of "Design languages you have used for data mining/analysis in the past year" in 2012, R was in the first place, defeating SQL and Java.
- Due to the popularity of R, database manufacturers such as Teradata and Oracle both released products supporting R.

**2) Excel (29.8 %) :**

- Excel, a core component of Microsoft Office, provides powerful data processing and statistical analysis capability, and aids decision making.
- When Excel is installed, some advanced plug-ins, such as Analysis ToolPak and Solver Add-in, with powerful functions for data analysis are integrated but such plug-ins can be used only if users enable them.
- Excel is also the only commercial software among the top five.

**3) Rapid-I Rapidminer (26.7 %) :**

- Rapidminer is an Open source software used for data mining, machine learning, and predictive analysis. In an investigation of KDnuggets in 2011, it was more frequently used than R.
- Data mining and machine learning programs provided by RapidMiner include Extract, Transform and Load (ETL), data pre-processing and visualization, modelling, evaluation, and deployment.
- The data mining flow is described in XML and displayed through a graphic user interface (GUI). RapidMiner is written in Java. It integrates the learner and evaluation method of Weka, and works with R.
- Functions of Rapidminer are implemented with connection of processes of Big Data Analysis operators. The entire flow is deemed as a production line of factory, with original data input and model results output.
- The operators can be regarded as specific functions and feature different input and output characteristics.

**4) KNIME (21.8 %) :**

- KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-source-rich data integration, data processing, data analysis, and data mining platform.
- It allows users to create data flows or data channels in a visualized manner, to selectively run some or all analytical procedures, and provides analytical results, models, and interactive views.
- KNIME was written in Java and, based on Eclipse, provides more functions as plug-ins. Through plugin files, users can insert processing modules to files, pictures, and time series, and integrate them into various Open source projects. **E.g. :** R and Weka.
- KNIME controls data integration, cleansing, conversion, filtering, statistics, mining, and finally data visualization. The entire development process is conducted under a visualized environment.

- KNIME is designed as a module-based and expandable framework. There is no dependence between its processing units and data containers, making them adaptive to the distributed environment and independent development.
- In addition, it is easy to expand KNIME. Developers can effortlessly expand various nodes and views of KNIME.

**5) <u>Weka/Pentaho (14.8 %) :</u>**

- Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java.
- Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc.
- Pentaho is one of the most popular open-source commercial intelligent software. It is a BI kit based on the Java platform.
- It includes a web server platform and several tools to support report, analysis, chart, data integration, and data mining, etc., all aspects of BI.
- Weka's data processing algorithms are also integrated in Pentaho and can be directly called.