gradeup

Big Data Systems

Content :-

Big Data :-Big Data is also **data** but with a **huge size**. It used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short this data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Example of Big Data :-

1. The **Stock Exchange** generates about *one terabyte* of new trade data per day.
2. **Social Media:-** The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, instagram every day. This data is mainly generated in terms of pictures and video uploads, putting comments, message exchanges etc.
3. A single Jet engine can generate 10+terabytes of data in almost 30 minutes of flight time. With thousands of flights per day, generation of data reaches up to many Petabytes.

Types Of Big Data:-

Big Data could be found in three forms:

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

Structured

Any data that can be accessed , stored and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiples of zettabytes.

Fact: 10^{21} bytes equal to 1 zettabyte or one billion terabytes forms a zettabyte**.**

> **Example of structured data:** Data stored in a relational database management system.

Examples Of Structured Data:

An 'Employee' table in a database is an example of Structured Data

EMP_ID	EMP_NAME	Branch	Gender	Salary	Grade	Project
001	Shruti	Management	F	5000	A	Evaluated
002	Ravi	HR	M	10000	A	Pending
003	Shaym	IT	M	9000	B	Pending
004	Ram	Management	M	20000	C	Evaluated
005	PK	IT	F	15000	C	Submitted
006	Rahul	IT	M	14000	A	Submitted
007	Raman	IT	M	21000	B	Evaluated

Unstructured

Any data with unknown form or there is no specific structure is classified as unstructured data. In addition the size of unstructureddata is huge, it poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Examples Of Un-structured Data

The output returned by 'Google Search'

Semi-structured

Semi-structured data can contain both the forms of data ie. Structured and un-structured. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

Examples Of Semi-structured Data

Personal data stored in an XML file-

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>

<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>

<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>

<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

Characteristics Of Big Data

(i) Volume – Asthe name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, volume of a data decides whether a particular data can actually be considered as a Big Data or not. Hence, 'Volume' is one of the characteristic which needs to be considered while dealing with Big Data.

(ii) Variety

Variety refers to the heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, databases and spreadsheets were the only sources of data considered by most of the applications. Nowadays, data in the form of photos,emails, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, analysing and mining data.

(iii) Velocity –'Velocity' refers to the speed of generation of data. How quickly the data is generated and processed to meet the demands, determines real potential in the data.
Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, mobile devices, social media sites, sensors, etc. The flow of data iscontinuous and massive.

(iv) Variability – 'Variability' refers to inconsistency which can be shown by the data at times, thus hampering the process of being able to manage and handle the data effectively.

Benefits of Big Data Processing

Benefits of big data processing includes :

- Before taking decisions businesses can utilize outside intelligence.

Access to social data from search engines and sites liketwitter, facebook are enabling organizations to fine tune their business strategies.

- Improvement in customer service

Traditional customer feedback systems are getting replaced by the new systems designed with Big Data technologies. In these new systems, natural language processingtechnologies and Big Data are being used to read and evaluate consumer responses.

- If there is any risk to the product/services then it was early identified
- Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

Distributed File System(DFS) :-It is a file system in which data stored on a server. The data is processed and accessed as if it was stored on the local client machine. The DFS makes it convenient to share information and files among the user on a network in authorized and controlled way. The server allows the client users to store data and share files just like they are storing information locally. However, the servers have full control over data and give access control to the clients.

Big Data Architecture:-

Photo -1

The 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.

Important Component Of Big Data Architecture:-

Data sources: The big data solutions start with one or more data sources. Examples include:

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

Data storage: The 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 known a data lake. Options for implementing this storage include blob containers in Azure Storage or Azure Data Lake Store.

Batch processing: Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to aggregate ,filter , and otherwise prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to the new files. Options include running U-SQL jobs in Azure Data Lake Analytics, using pig, Hive, or custom Map/Reduce jobs in an HDInsight Hadoop cluster, or using python ,Java or Scala programs in an HDInsight Spark cluster.

Real-time message ingestion: If the solution includes real-time sources, then the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where the incoming messages are dropped into a folder for processing. However, many of the solutions which need the 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, Kafka and Azure IoT Hubs.

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

Analytical data store: Many big data solutions prepare the 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 that can be a Kimball-style relational data warehouse, as seen in most traditional business intelligence (BI) solutions. Alternatively, data could be presented through the low-latency NoSQL technology such as the 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 the large-scale, cloud-based data warehousing. HDInsight supports Interactive HBase,Hive, and Spark SQL, which can also be used to serve data for analysis.

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

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

gradeup SUPER

Gradeup UGC NET
Super Superscription

Features:

1. 7+ Structured Courses for UGC NET Exam
2. 200+ Mock Tests for UGC NET & MHSET Exams
3. Separate Batches in Hindi & English
4. Mock Tests are available in Hindi & English
5. Available on Mobile & Desktop

Gradeup Super Subscription, Enroll Now

gradeup.co