**Name: Jigar Siddhpura SAPID:** 60004200155

**DIV: C/C2 Branch:** Computer Engineering

BDI - Experiment 1

**Aim** : Case study on Big Data technologies used in ISRO

**Characteristics**

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many multinational companies to process the data and business of many organizations. The data flow would exceed 150 exabytes per day before replication.

There are **five v's of Big Data** that explains the characteristics.

1. **Volume**: the size and amounts of big data that companies manage and analyze
2. **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 quantifiable business benefits
3. **Variety**: the diversity and range of different data types, including unstructured data, semi-structured data and raw data
4. **Velocity**: the speed at which 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
5. **Veracity**: the “truth” or accuracy of data and information assets, which often determines executive-level confidence

Organization

The Indian Space Research Organisation (ISRO) is the national space agency of India, renowned for its remarkable achievements in space exploration and technological development. Founded in 1969, ISRO operates under the Department of Space and boasts a track record of self-reliance in building satellites, launch vehicles, and ground stations. They spearhead a diverse range of projects, from Earth observation for agriculture and disaster management to lunar and Mars missions, demonstrating India's growing presence in the global space arena. ISRO's contributions extend beyond scientific pursuits, impacting socio-economic development through innovations in telemedicine, weather forecasting, and communication technologies. This vibrant institution continues to push the boundaries of space exploration, inspiring future generations to reach for the stars.

**Big Data tecnologies used** :

* **Data Storage** -

**1. Apache Hadoop:**

The Indian Remote Sensing Data Archive (IRSDA) leverages Hadoop's prowess to manage Earth observation data, enabling scientists to study land cover changes, deforestation, and crop patterns.

**2. MongoDB:**

 The Space Applications Centre (SAC) uses MongoDB to store telemetry data from satellites, allowing real-time monitoring of spacecraft health and performance. Additionally, the Indian Institute of Remote Sensing (IIRS) utilizes it for geospatial data storage and analysis in disaster management and environmental monitoring.

**3. RainStor:**

 The Indian Astronomical Observatory (IAO) utilizes RainStor to store and analyze massive datasets from telescopes, enabling astrophysicists to study celestial objects and phenomena with greater precision.

**4. Hunk:**

 ISRO utilizes Hunk to archive historical mission data and scientific observations, ensuring its preservation for future research and analysis.

**5. Cassandra:**

 While specific examples within ISRO are less documented, Cassandra's capabilities make it a potential candidate for real-time data processing and analysis scenarios.

* **Data Mining -**

**1. Presto:**

Fast & Interactive Data Exploration: Presto's lightning-fast querying empowers ISRO scientists to interactively analyze massive datasets without the need for traditional data warehousing. Examples: Analyzing Earth observation data for real-time flood monitoring, studying satellite telemetry for anomaly detection, and exploring vast astronomical datasets for hidden patterns.

**2. RapidMiner:**

ISRO used this to predict crop yields based on satellite imagery, classifying celestial objects from telescope data, and identifying anomalous spacecraft behavior for potential maintenance needs.

**3. Elasticsearch:**

It excels at indexing and searching large volumes of data, enabling efficient retrieval and analysis of real-time information. It enables scientists to quickly search through telescope observations for specific objects, monitoring satellite health data in real-time for anomalies, and providing fast access to archived weather data for climate studies.

* **Data Analytics -**

1. **Apache Kafka** -

It as a high-speed data highway! Kafka efficiently streams real-time data from satellites, telescopes, and ground stations, ensuring continuous flow for immediate analyssis. Examples: Monitoring spacecraft health parameters in real-time, tracking weather patterns for immediate flood prediction, and analyzing telescope data for real-time anomaly detection.

1. **Splunk** -

It acts like a data detective, searching through massive datasets from various sources to identify patterns and anomalies. Think mission logs, sensor readings, and even social media data. Examples: Identifying potential equipment failures on satellites before they occur, analyzing social media trends to gauge public perception of space missions, and investigating cyber security threats in ISRO's IT infrastructure.

1. **KNIME** -

It is a visual analytics platform that allows scientists to build complex data analysis workflows without writing a single line of code. Examples: Building models to predict crop yields based on satellite imagery and weather data, classifying celestial objects from telescope data, and creating interactive dashboards to visualize spacecraft health parameters.

**Ecosystem Components**

ISRO uses Hadoop as a core component of its *major* big data approach. Here are the components used by it:

* **Hadoop Distributed File System (HDFS):** HDFS is a distributed file system that enables ISRO to store and manage large volumes of structured and unstructured data across multiple nodes in a scalable and fault-tolerant manner. ISRO uses HDFS as a core component of its data storage layer.
* **Apache Spark:** Spark is a distributed computing framework that runs on top of Hadoop and enables ISRO to process and analyze large volumes of data in a fast and efficient manner. ISRO uses Spark for various data processing tasks, such as data cleaning, transformation, and machine learning.
* **Apache Superset:** Superset is an open-source data visualization platform that runs on top of Hadoop and enables ISRO to create interactive dashboards and reports that provide insights into its data. ISRO uses Superset to visualize its data in a variety of ways, such as charts, graphs, and maps.

**Advantages**

* Scalability: Hadoop ecosystem is designed to scale horizontally by adding more nodes to the cluster, making it a good choice for storing and processing large volumes of data.
* Fault-tolerance: Hadoop's distributed architecture provides fault-tolerance by replicating data across multiple nodes, ensuring that data is not lost in case of node failures.
* Flexibility: The Hadoop ecosystem is modular and allows businesses to choose from a wide range of tools and technologies to meet their specific needs.
* Cost-effective: Hadoop is open-source software, and many of the components in the Hadoop ecosystem are also open-source, which can significantly reduce the cost of implementing a big data approach.

**Disadvantages**

* Complexity: The Hadoop ecosystem is complex and requires specialized skills to manage and maintain the infrastructure, which can be a significant investment for businesses.
* Latency: Hadoop's batch processing approach may not be suitable for real-time applications that require low-latency processing.
* Data Security: Hadoop's distributed architecture can make it more challenging to manage data security, as data is spread across multiple nodes in the cluster.
* Compatibility: Some legacy systems may not be compatible with Hadoop, which can create integration challenges.