Satellite Data Pipelining System for Space Agencies

Kundan Thakur

What is it actually?

A large-scale data pipeline designed to process, store, and deliver satellite data ranging from terabytes to petabytes. It ensures scalability, high availability, and fault tolerance for mission-critical space applications.

How will it work?

"Terabytes to petabytes," "real-time," and etc. immediately scream distributed systems and big data architecture. I'll need a layered approach: ingestion, raw storage, processing, processed storage, and finally access. Message queues, distributed file systems, and computation frameworks will be central.

Key components

1. **Raw Data Storage Layer:** The source of all kinds of raw imagery, and other sensor data. Must be stored and secured in raw and untouched format. This must be immutable.
2. **Data Ingestion Layer:** As the name suggest, it will integrate the incoming data streams into multiple ingestion nodes, which will help in load balancing.
3. **Ingestion API**: Tools designed to receive data, perform basic validation, and push it to a message queue.
4. **Message Queue (e.g., Kafka)**: A high-throughput, fault-tolerant buffer that decouples the ingestion from processing. It will be essential for real-time streams and handling backpressure.

Key components

5. **Processing & Analytics Layer**: This layer will be responsible for processesing data from the message queue in near real-time (e.g., filtering, real-time alerts). Runs complex, long-running analyses on vast amounts of raw data in the distributed storage.

6. **Scalable Processed Storage Layer:** Stores derived datasets, aggregated results, and mission-specific products. Optimized for fast retrieval and analytical queries.

7. **Access & Retrieval Layer**:RESTful APIs : Provides structured access to processed data for scientists and applications. Allows direct, interactive querying of data in the processed storage. Tools for end-users to explore and visualize data. Ensures only authorized users can access specific datasets, often integrated with an Identity Management system.