SYSTEM DESIGN CHALLENGE

Goal:

The goal is to design a system that processes data from various upstream services , published to an event broker. The data needs to be made available to employees for insights and dashboards.

Approach to the Challenge

The approach is divided in 3 steps to get initial insights and ensure regular evaluation of the approach so that it can be adjusted and reworked as per the requirements.

1. Stakeholder Discussions:
   Conduct discussions with stakeholders to understand :
   - Business Goals: The strategic objective that the data solution aims to achieve
   - Key Metrics: Target KPIs for the business
   - Data Sources: What are the data sources? Understand the nature , data quality and format of data generated by each upstream service.
   - Data Usage: How different teams use the data? What insights are of relevance to each stakeholder
   - Frequency and Volume: Expected frequency of data generation and volume of data
2. Technology Evaluation:
   - Data Ingestion: Evaluate tools based on batch or stream processing .
   - Data Storage : Consider scalability , who would access the data and then based on that decision is made regarding a OLAP or OLTP system
   - Data Processing : Asses the expected data volume and consider if parallel processing frameworks such as Apache Spark are required or not.
   - Cloud or On Premise : Cloud centric approach should be taken or on premise.
   - BI Tools: Identify user friendly tools like Tableau or Power BI for data visualisation
   - Tools for Data Quality and Governance

3. Proof of Concept (PoC):
   - Pilot Implementation: Implement a pilot with subset of data to validate chosen technology and architecture
   - Testing: Conduct performance,user and load testing to ensure system meets scalability and performance requirements

4. Full Scale Implementation:
    - Incremental Rollout: Gradual scale system , ensuring minimal disruption
    - Continous Monitoring

Data Architecture :

Assumptions:

- Frequency of pipeline: The pipeline runs in batch mode and is executed multiple times a day (e.g., every hour) to provide near real-time insights
- Data Volume: Lets assume there are 15 Upstream Services sending data to broker with each creating 10gb data everyday. Since the pipeline runs daily so:
  Daily Data Volume : 10Gb*15Services = 150GB/day
  Monthly Data Volume : 15Gb/day *30 days = 4.5TB/Month
- On Premise or Cloud Solution: Cloud
- Users: Business Analysts, Data Scientists across the company for insights and Dashboards
- Data Type: Data includes structured,semi-structured and unstructed formats.

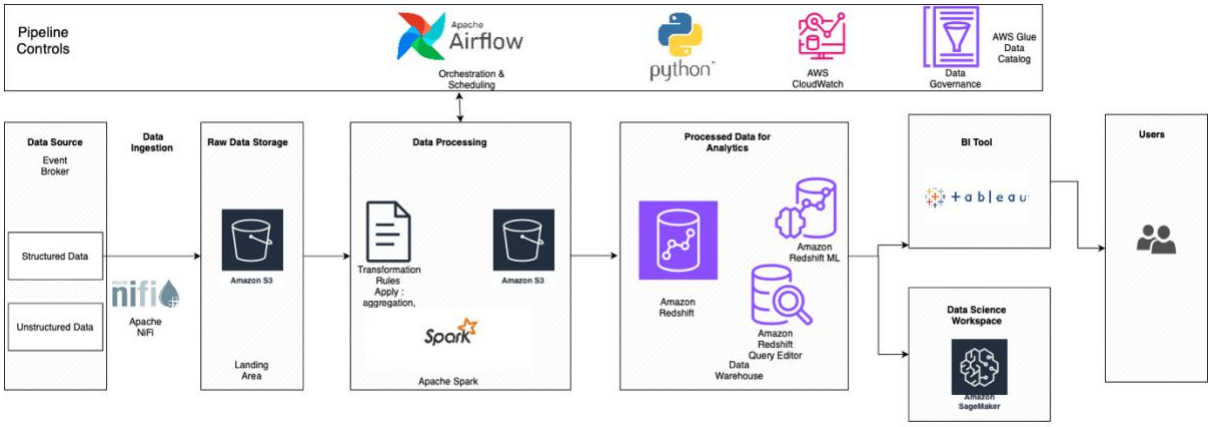

Figure : Data Architecture

Data Ingestion:

Tool: Apache Nifi

Reasons : Ingests data from event broker in a user friendly interface for data flows.

It can handle variety of data formats and sources and can handle high frequency data efficiently so in case the system scales then can handle the updated data volume.

Challenges:

It introduces additional processing overhead and can also be resource intensive.

Raw Data Storage

Tool: AWS S3

Reasons: It is highly scalable with leveraging various storage classes and having a strategy for moving data between them. It makes it also economical for storing large datasets.

Challenges: Latency can be an issue in comparison to a database

Data Processing

Tool: Apache Spark and AWS S3

Reasons : Efficiently processes large datasets in distributed manner. High performance for handling large datasets. Also python has a good support for Apache Spark – pyspark library

Challenges:

It requires expertise to optimize and configure the resource management. Also processing is resource intensive as it has in-memory computing.

Data Warehouse

Tool: Amazon Redshift

Reasons : Omptimized for complex queries and analytical workload

Challenge: It has high cost compared to traditional databases.

Data Governance and Orchestration:

Tool: Apache Airflow, AWS Glue Catalog, AWS Cloudwatch

Reasons : Efficiently schedules and manages the workflow . AWS Cloudwatch has comprehensive monitoring and alterting .

AWS Glue Catalog has a metadata repository

Challenge : Operational Overhead

Analytics and Visualisation

Tools: Tableau, AWS SageMaker

Reasons : Tableau provides very intutive data visualisation capability and SageMaker has good machine learning capabilities.

Challenge: Licensing and usage costs involved.

Technological And Organisational Challenges

Technical Challenges

1. Data Volume and Velocity;
   Challenge : Manage large volumes of data .
   Mitigation : Use partitioning strategies
2. Data Quality:
   Challenge: Ensure data quality and consistency
   Mitigation: Implement data validaton,cleansing and take test driven TTD approach in ETL Jobs
3. Scalability:
   Challenge: Ensuring the system can scale with increasing data volume and user load.
   Mitigation: Use cloud-based services like Amazon S3 and Redshift that automatically scale and leverage distributed processing with Apache Spark.

Organisational Challenges

1. User training:
   Challenge: Ensuring all employees are proficient in using BI tools and understanding the data.
   Mitigation: Provide comprehensive training programs and documentation. Conduct workshops and create user guides.
2. Collaboration:
   Challenge: Facilitating collaboration between data engineers, analysts, and business users.
   Mitigation: Set up regular meetings, use collaborative tools like Jira and Confluence, and establish clear communication channels.
3. Data Governance:
   Challenge: Implementing data governance and access control.
   Mitigation: Define data governance policies, use AWS IAM for access control, and implement auditing and logging mechanisms.