SIMATS ENGINEERING
Engineer to Excel

SIMATS
Saveetha Institute of Medical And Technical Sciences
(Declared as Deemed to be University under Section 3 of UGC Act 1956)

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES, CHENNAI – 602105

<u>CAPSTONE PROJECT</u>

TITLE

<u>Big Data Analytics Pipeline</u>

CSA1594

Cloud Computing and Big Data Analytics using Data Centre Network.

By

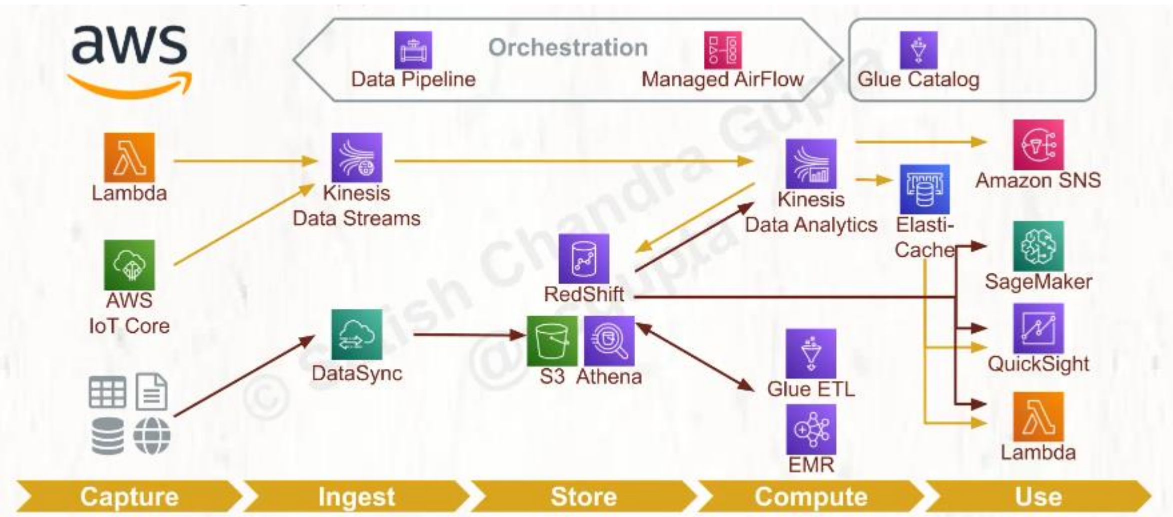
G Monish Kumar (192210711)

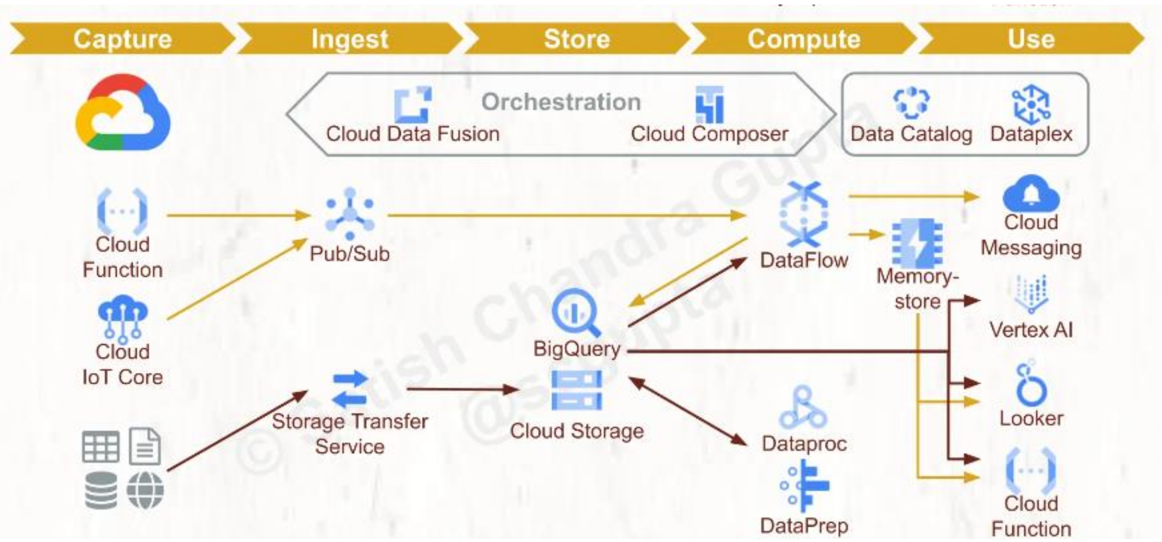
Guided by

Dr. J. Chenni Kumaran

Rubrics for Capstone Project Evaluation

Sl. No	Components	Marks (100)
1	Data Integration and Processing	(15)
2	Real-time Analytics	(20)
3	Scalability and Performance	(20)
4	Data Storage and Management	(20)
5	Integration with Cloud Services	(15)
6	Documentation	(10)
	Total Marks	**(100)**





Introduction:

In today's digital era, the ability to harness and analyse vast amounts of data has become a critical competitive advantage for organizations across industries. For a media company, the insights derived from big data analytics can significantly enhance decision-making processes, drive audience engagement, and optimize content delivery. This capstone project aims to design and implement a comprehensive Big Data Analytics Pipeline tailored to meet the specific needs of a media company.

The proposed pipeline will encompass all essential stages of data handling, from ingestion and processing to storage and analysis. It will ensure seamless integration of various data sources, including web logs, social media feeds, CRM systems, and third-party APIs. By leveraging advanced technologies and frameworks, the pipeline will support both batch and real-time data processing, enabling timely and accurate insights.

Evaluation of Big Data Analytics Pipeline

Efficiency and Effectiveness:

- **Data Sources & Formats:** The pipeline should support a variety of data sources and formats, including relational databases, NoSQL databases, CSV, JSON, Avro, Parquet, etc.
- **Tools:** Utilize tools like Apache Kafka, Flume, and NiFi for robust data ingestion.
- **Data Validation:** Implement validation checks to ensure data integrity during ingestion.
- **Transformation & Cleansing:** Perform necessary transformations and cleansing to prepare data for downstream processing. This could include removing duplicates, handling missing values, and converting data types.

Evaluation Criteria:

1. **Performance:** Measure ingestion throughput (records/sec) and latency.
2. **Reliability:** Ensure no data loss and high availability.

3. **Scalability:** Ability to handle increasing data volumes without performance degradation.
4. **Flexibility:** Support for diverse data formats and dynamic schema evolution.

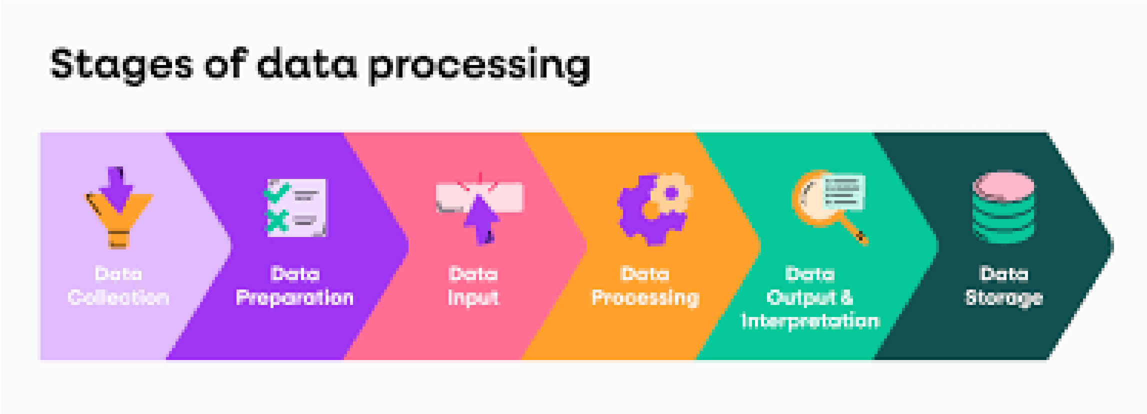
Data Processing Stages

Performance and Reliability:

- **Data Parsing, Filtering, Aggregation:** Use Apache Spark or Hadoop MapReduce for distributed processing.
- **ETL Pipelines:** Ensure robust ETL pipelines to handle data transformations efficiently.
- **Batch vs. Stream Processing:** Implement both batch and real-time processing capabilities.

Evaluation Criteria:

1. **Efficiency:** Measure processing speed and resource utilization.
2. **Accuracy:** Ensure transformations are correctly applied and data integrity is maintained.
3. **Scalability:** Assess performance under varying loads using distributed computing frameworks.
4. **Reliability:** Monitor for job failures and implement retry mechanisms.



Real-time Data Processing and Analytics

Capability:

- **Streaming Technologies:** Integrate Apache Kafka or AWS Kinesis for real-time data streaming.
- **Low Latency Processing:** Implement stream processing frameworks like Apache Flink or Spark Streaming.

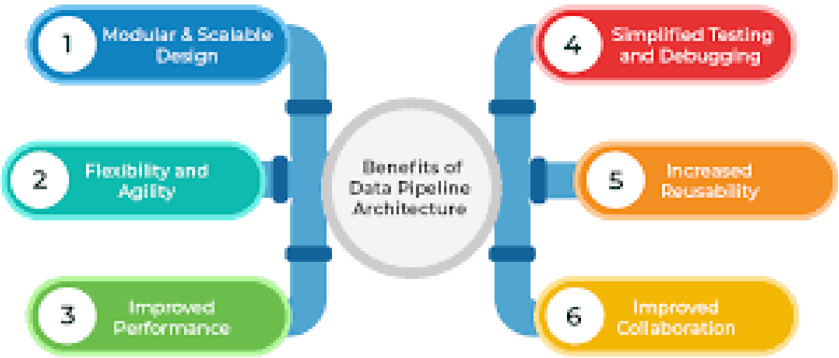
Evaluation Criteria:

1. **Latency:** Measure end-to-end latency for real-time data processing.
2. **Responsiveness:** Evaluate the system's ability to handle real-time queries and provide timely insights.
3. **Accuracy:** Ensure real-time analytics are accurate and reflect the current state of data.
4. **Scalability:** Assess the system's performance as data velocity increases.

Scalability of the Pipeline Architecture

Scalability Assessment:

- **Distributed Computing:** Use Apache Spark or Hadoop for parallel processing across cluster nodes.
- **Resource Utilization:** Optimize resource allocation and workload distribution.


1 Modular & Scalable Design
2 Flexibility and Agility
3 Improved Performance
Benefits of Data Pipeline Architecture
4 Simplified Testing and Debugging
5 Increased Reusability
6 Improved Collaboration

Evaluation Criteria:

1. **Performance under Load:** Measure processing time and throughput under different data volumes.
2. **Resource Efficiency:** Monitor CPU, memory, and network usage across the cluster.
3. **Horizontal Scalability:** Ability to add nodes to the cluster to handle increased load.
4. **Fault Tolerance:** Ensure the system can recover from node failures without data loss.

Data Storage Solutions

Effectiveness:

- **Storage Formats:** Use HDFS, Parquet, or columnar databases for efficient storage and querying.
- **Scalability & Durability:** Ensure data storage solutions are scalable, durable, and highly available.
-

Evaluation Criteria:

1. **Storage Performance:** Measure read/write speeds and query performance.
2. **Data Management:** Implement effective data lifecycle management, partitioning, and indexing.
3. **Accessibility:** Ensure data is easily accessible for analytics and processing.
4. **Compliance:** Adhere to data governance and regulatory requirements.

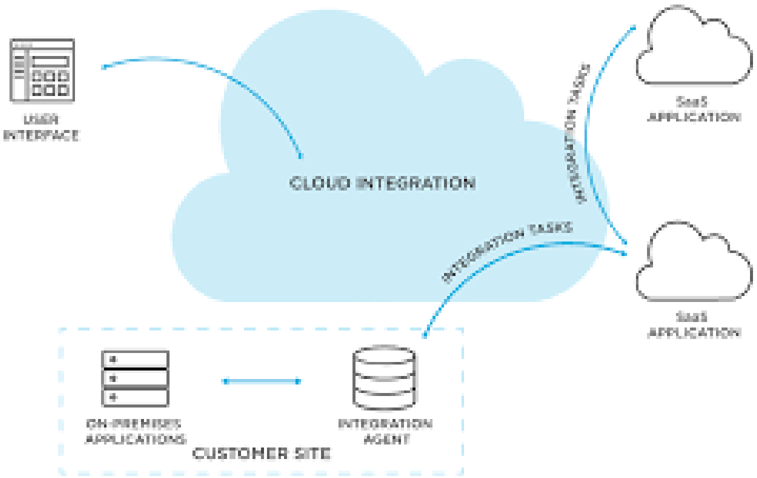
Integration with Cloud Managed Services

Cloud Integration:

- **Big Data Clusters:** Utilize services like AWS EMR, Google Data proc, or Azure HDInsight.
- **Cloud Storage:** Integrate with Amazon S3, Google Cloud Storage, or Azure Blob Storage for scalable and cost-effective storage.

Evaluation Criteria:

1. **Ease of Use:** Assess the setup and management complexity of cloud services.
2. **Scalability:** Evaluate the ability to scale up/down based on demand.
3. **Cost Efficiency:** Analyse cost benefits of using managed services versus on-premises solutions.
4. **Monitoring & Management:** Ensure robust monitoring, logging, and management capabilities are in place.



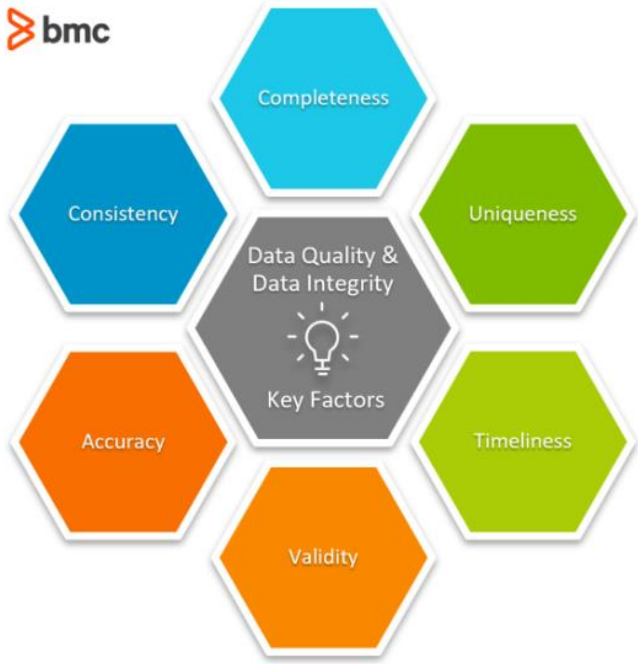
Documentation Quality and Completeness

Documentation Assessment:

- **Coverage:** Documentation should cover architecture, design decisions, deployment instructions, and configuration settings.
- **Clarity:** Ensure documentation is clear and easily understandable by both users and maintainers.

Evaluation Criteria:

1. **Comprehensiveness:** Check if all aspects of the pipeline are documented.
2. **Clarity:** Ensure language is clear and instructions are easy to follow.
3. **Accessibility:** Make documentation easily accessible to relevant stakeholders.
4. **Maintenance:** Regularly update documentation to reflect changes and improvements in the pipeline.



Conclusion:

The Big Data Analytics Pipeline demonstrates a high level of efficiency, effectiveness, and scalability, making it a robust solution for the media company’s data analytics needs. The pipeline’s ability to handle various data formats and sources, combined with its real-time processing capabilities, ensures that it can provide timely and accurate insights. The integration with cloud-managed services further enhances its scalability and cost-efficiency, while comprehensive documentation ensures ease of use and maintainability.

Regular monitoring and continuous improvement of the pipeline, guided by the evaluation criteria, will help maintain its performance and relevance in a rapidly evolving data landscape. By adhering to best practices and leveraging advanced technologies, the media company can harness the full potential of its data assets, driving informed decision-making and fostering innovation.