Big Data Processing

- Lecturer: Mr. CHAN Sophal
- Presented by: TOUCH Sopheak
  HENG Seyha

Content

- Problem Statement
- Tasks
- Proposed Solution
- Conclusion

Problem Statement

- AMS Tech, a medium-sized fintech company, has recently expanded its services, resulting in a significant increase in data volume.
- The data comes from various sources such as transaction logs, customer interactions on their web and mobile platforms, and third-party financial services.

Problem Statement (Cont.)

- This data is crucial for AMS Tech to analyze user behavior, detect fraud, and tailor their services.
- However, the current data infrastructure is struggling to keep up with the scale and diversity of data, which includes structured data from transaction logs and unstructured data from user interactions.

Problem Statement (Cont.)

- AMS Tech's existing data storage solutions, primarily traditional SQL databases like PostgreSQL, are not able to efficiently handle the current scale of data, especially the unstructured part.
- The data ingestion process is slow and often creates bottlenecks, hindering real-time analysis and timely decision-making.

Problem Statement (Cont.)

- Furthermore, the company wants to leverage technologies like Elasticsearch for enhanced data search and analytics capabilities but is unsure how to integrate it effectively with their current system.

Tasks

- As a data architect, we are asked to redesign AMS Tech's data ingestion and storage architecture to address these challenges.
- Our proposal will cover the following aspects:

1. Data Ingestion Strategy
2. Data Storage Solution
3. Integration and Scalability
4. Performance and Efficiency
5. Security and Compliance Considerations

Proposed Solution

Data Ingestion Strategy:

- Recommendation: Apache Kafka for Real-time Stream Processing

Proposed Solution (Cont.)

Data Storage Solution:

- Proposed Solution: Apache Hadoop Distributed File System (HDFS) with Elasticsearch for Search and Analytics

Proposed Solution (Cont.)

Integration and Scalability:

- Integration:

- Kafka can be seamlessly integrated with HDFS through connectors, ensuring a smooth flow of data from ingestion to storage.

- Tools like Apache NiFi can be employed for orchestrating data flows and managing the integration of different storage components.

Proposed Solution (Cont.)

Integration and Scalability: (Cont.)

- Scalability:

- Kafka and HDFS are horizontally scalable, allowing AMS Tech to scale their infrastructure with growing data volumes.

- Elasticsearch provides scalability by adding nodes to the cluster, accommodating increased search and analytics demands.

Proposed Solution (Cont.)

Performance and Efficiency:

- Expected Performance Improvements: Ingestion Speed, Data Processing, Query Response Times
- Potential Efficiencies: Cost Savings and Resource Utilization

Proposed Solution (Cont.)

Security and Compliance Considerations:

- Security Measures

- Implement end-to-end encryption for data in transit using technologies like SSL/TLS.

- Enforce access controls and authentication mechanisms at various layers (Kafka, HDFS, Elasticsearch).

- Regularly audit and monitor data access for suspicious activities.

Proposed Solution (Cont.)

Security and Compliance Considerations: (Cont.)

- Compliance:

- Ensure compliance with financial data regulations, such as PCI DSS or GDPR, by implementing necessary controls.

- Regularly audit and document security practices to meet regulatory requirements.

Conclusion

- This proposed architecture aims to address AMS Tech's challenges by leveraging **robust** and **scalable** technologies for data ingestion, storage, and analytics.
- It ensures real-time processing, scalability, and efficient handling of diverse data types while considering security and compliance aspects.