Data Warehouse Architecture Proposal Based on AWS services

Mikhail Lysikov

Task

- Propose an architecture for a new data warehouse

Input Data

- Various data sources
  - 5 CRM (million rows)
  - MySQL database (million rows)
  - Clickhouse database (billion rows)
  - Google Drive files (Excel) (million rows)
- Different types of data
  - Structured
  - Unstructured

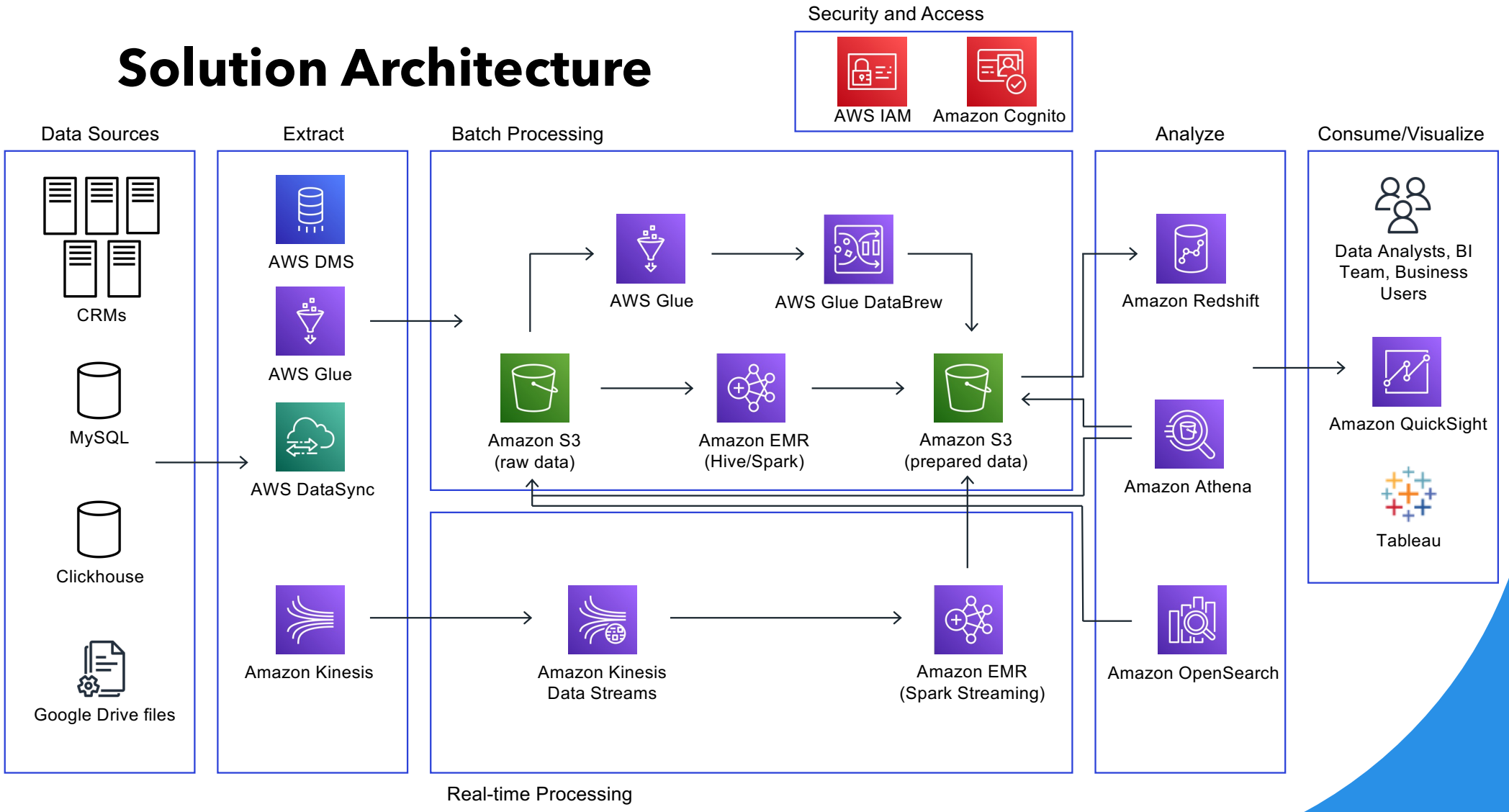
Requirements

- Store current and historical data
- Implement data masking to hide sensitive data
- Ensure data quality
- Provide daily and real-time reporting
- Apply the four V's of Big Data

Preliminary Estimates

- Several billion rows in a data lake
- Hundreds of millions of rows in a data warehouse
- Daily increment - tens of millions of rows
- If we assume that the average row size is 500 bytes, then we get
  - Data lake size – 931 GB (raw data) + 400-500 GB (prepared data)
  - Data warehouse size – 232 GB
  - Daily increment – 23 GB

Solution Architecture

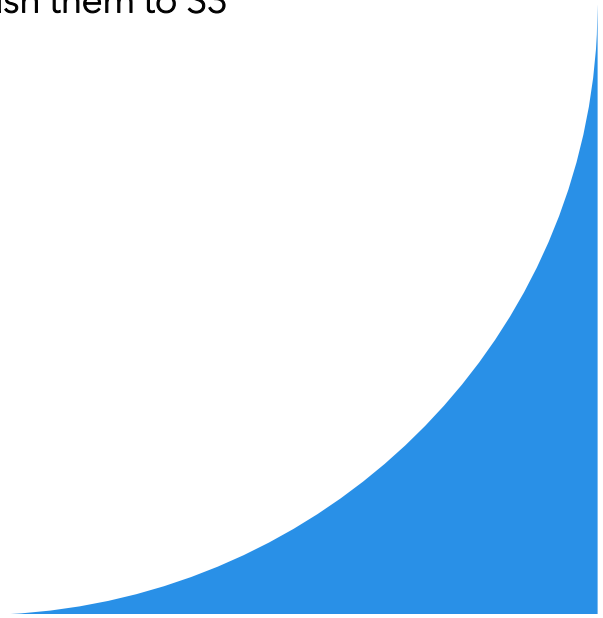


Architecture Explained

- Data warehouse on AWS
- Data Lakehouse and Lambda approaches have been chosen for the current architecture
- The proposed architecture allows
  - Integrate data from various data sources
  - Store all data in one place
  - Handle structured/unstructured data
  - Clean, validate, and transform big data
  - Process data in batches and in real-time
  - Explore data, conduct ad-hoc research and build reports

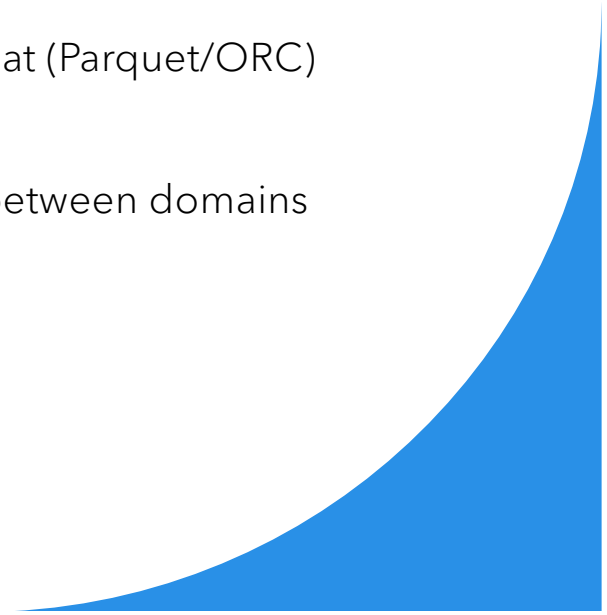
Extract Explained

- DMS to extract data from relational databases
- Glue jobs with appropriate JDBC drivers and PySpark logic for other databases
- Develop a Glue ETL job with appropriate algorithms to mask data
- There are several options to copy files from Google Drive:
  - Create a virtual server in Google Account, pull files from Google Drive, and push them to S3
  - Start up an EC2 instance and use the appropriate CLI commands to copy files
- DataSync for migration unstructured data from on-premises to AWS
- Kinesis to consume data in real-time



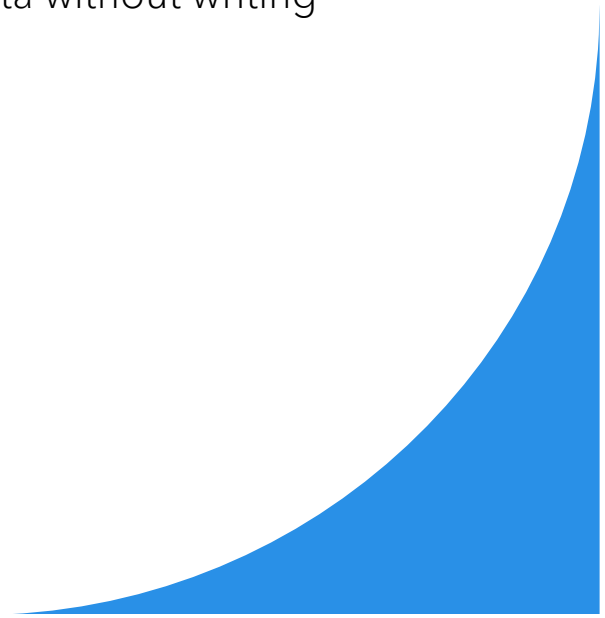
Storage Explained

- S3 as the primary storage platform
- Two data storage tiers
  - Raw
  - Prepared
- Raw data contains data in its original form and represents a single source of truth
- Prepared data contains cleaned, validated, and transformed data in format (Parquet/ORC) optimized for analytics
- (optional) The third level (domain level) can be introduced to split data between domains and cover the needs of interested teams



Batch Processing Explained

- Glue, Glue DataBrew, and EMR to clean, validate, enrich and transform data
- Orchestrate ETL jobs using AWS Step Functions and AWS Lambda
- For cost-effective and performance of data transformation, it makes sense to use an EMR cluster
- (optional) Glue DataBrew can be used to visually clean and normalize data without writing code



Real-time Processing Explained

- Depending on requirements, several processing options can be considered
  - Kinesis Firehouse -> S3 or Redshift
  - Kinesis Data Streams -> Spark Streaming on EMR -> S3 (prepared) (illustrated in the diagram)
  - Kinesis Data Streams -> Flink on EMR -> S3 (prepared)

Analyze Explained

- It would be appropriate to use Redshift for operational reporting due to a large amount of data
- Athena can help to analyze data in S3 (raw and prepared) using SQL to understand what data is available and how it can be used for future reporting
- (optional) If there is a need to analyze large volumes of unstructured text data, and perform a full-text search, then it would be appropriate to use Amazon OpenSearch

Consume Explained

- As a data visualization tool, either QuickSight or Tableau can be used
- QuickSight and Tableau have many connectors by which we can visualize data located in Redshift or S3 through Athena

Security and Access

- Depending on requirements
  - Configure S3 permissions using IAM and S3 bucker policies
  - Provide authentication, authorization, and user management using Amazon Cognito

Aligning with the Four V of Big Data

- Volume. Based on preliminary estimates, the initial size of the data warehouse is already almost two terabytes. S3 as the main storage platform allows to store and protect any amount of data
- Variety. Data comes from a wide variety of data sources (CRMs, files, transactional and analytical systems) in different formats (structured, unstructured)
- Veracity. ETL tasks using Glue, Glue DataBrew, and EMR help us solve the problems of data cleaning and validation
- Velocity. Our solution ensures the collection, storage, processing, and analysis of data in a fairly short time: from one day (batch layer) up to real-time (real-time layer)