Big Data Processing on Clouds

Tessema Mengistu

mengistu@vt.edu

Big Data Services on Clouds

- Data pipeline
  - An infrastructure that supports data-driven decision
  - It basically involves:
    - Collection
      - Cleansing
    - Storage and processing
    - Make decisions based on the result

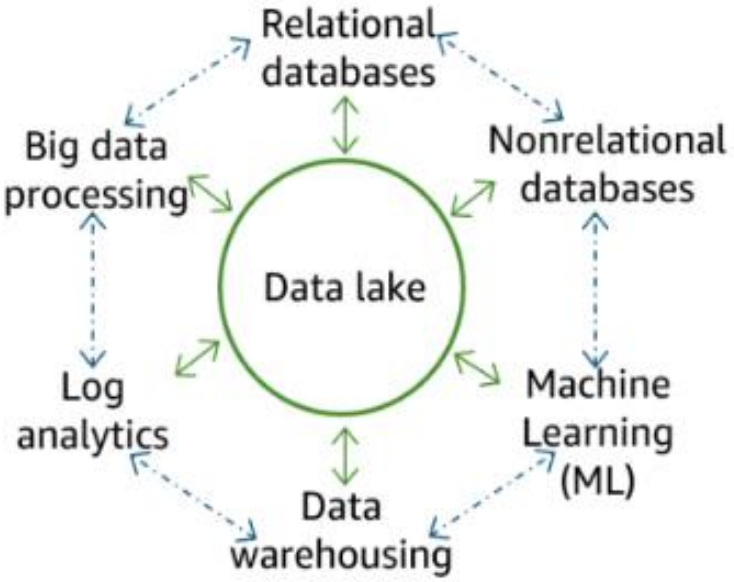
Big Data Services on Clouds

- Data Sources
  - Relational Databases
  - NoSQL
    - DynamoDB
      - A fully managed, scalable NoSQL database
      - A highly available key-value storage system
      - Supports both document and key-value store models and has been used for mobile, web, gaming, IoT, advertising, real-time analytics, and other applications
  - Applications
  - IoT Devices
  - Files
  - . . .

Big Data Services on Clouds

- Data storage plays a critical role in the performance of any big data processing
- Key points in designing modern data architecture
  - Scalability
  - Simpler data movement
  - Unified governance
  - Cost

Big Data Services on Clouds



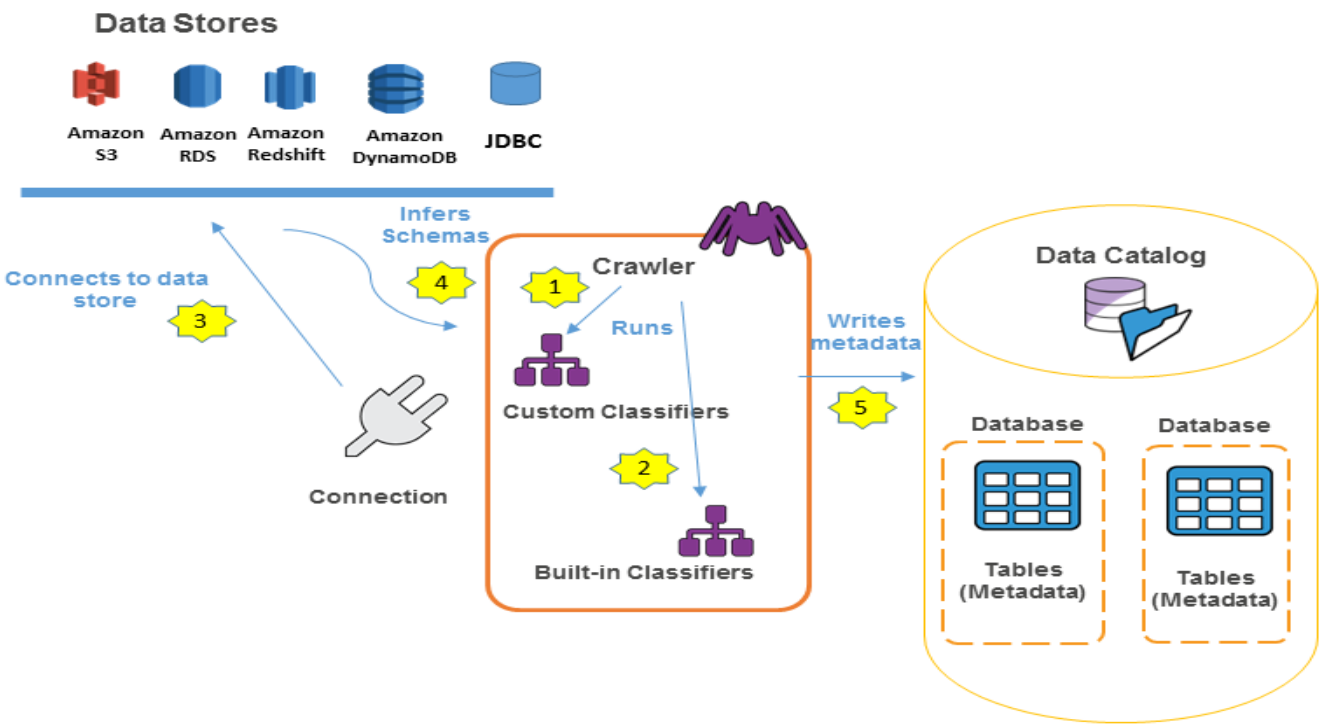
AWS Big Data Services

- AWS application data Integration - Ingestion
  - App Flow
    - Ingest data from applications
  - DMS – Database Migration Service
    - Ingest data from relational databases
  - DataSync
    - Ingest data from file systems
  - Data Exchange
    - Integrate data from a third-party data source

AWS Big Data Services

- AWS Glue
  - A serverless data integration service that is simple to use and is based on the Apache Spark engine
  - Enables to discover, analyze, and transform the data through Spark-based in-memory processing
  - **Glue crawlers** help autodetect the schema of source datasets and create virtual tables in **Glue Data Catalog**
  - Fully manages service

AWS Big Data Services



https://docs.aws.amazon.com/glue/latest/dg/catalog-and-crawler.html

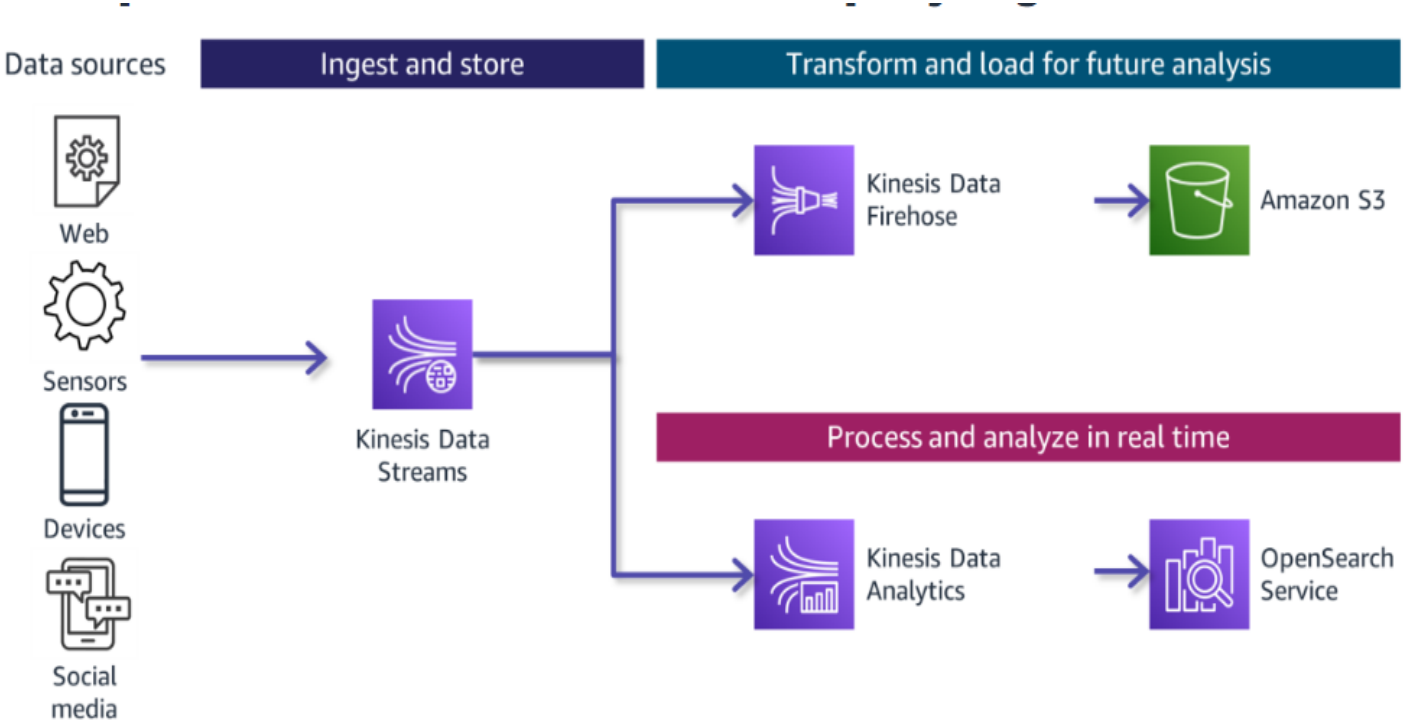
AWS Big Data Services

- Amazon Kinesis Data Streams (KDS)
  - Used to build real-time streaming pipelines for use cases such as website clickstreams, application log streams, and Internet of Things (IoT) device event streams
  - It provides:
    - Kinesis Producer Library (KPL)
      - Data producers can integrate to push data to Kinesis
    - Kinesis Consumer Library (KCL)
      - Data-consuming applications can integrate to access the data

AWS Big Data Services

- Amazon Kinesis currently offers four services:
  - Kinesis Data Firehose
    - Ingest and deliver streaming data
  - Kinesis Data Analytics
    - Performs analysis on streaming data
  - Kinesis Data Streams
  - Kinesis Video Streams

AWS Big Data Services



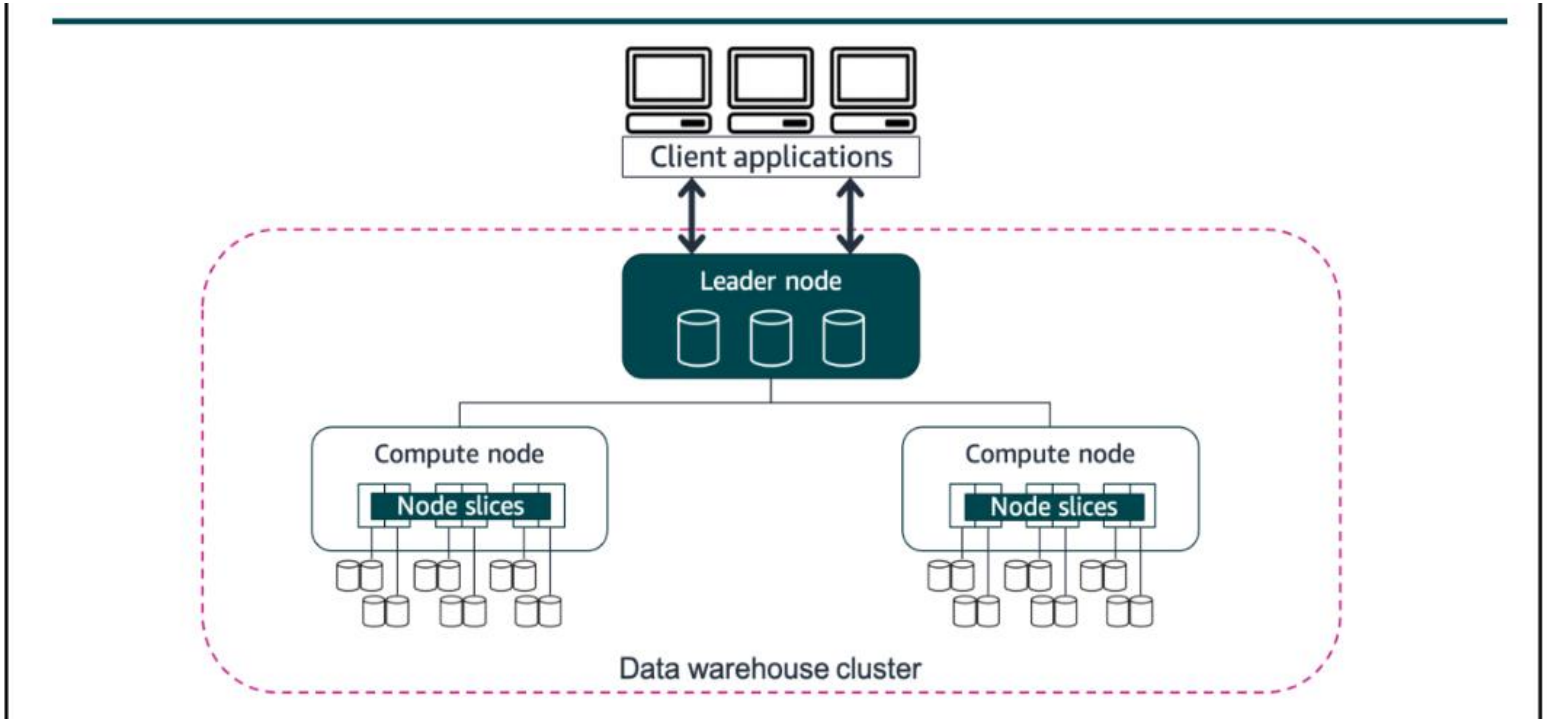
AWS Big Data Services

- Amazon Athena
  - An interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL
  - Fully managed and serverless
  - Built on open-source Trino and Presto engines and Apache Spark frameworks

AWS Big Data Services

- Amazon Redshift
  - Fully managed data warehouse service
  - Consists of a collection of resources called **nodes**
    - Organized into clusters
      - Leader node and compute nodes
      - Parallel processing
    - Runs Redshift engine
    - Based on PostgreSQL
      - Contains one or more column-oriented databases

AWS Big Data Services



AWS Big Data Services

- Multiple deployment options on:
  - Amazon Elastic Compute Cloud (EC2)
  - Amazon Elastic Kubernetes Service (EKS)
  - AWS Outposts

AWS Big Data Services

- AWS big data processing frameworks:

Batch Processing	Stream Processing	
Apache Spark	Amazon Kinesis	
Apache Hadoop MapReduce	Apache Spark Streaming	AWS Lambda
Apache Hive	Apache Hive	Apache Flink
Apache Pig	Apache Pig	Apache Storm

AWS Big Data Services

- Amazon EMR
  - An AWS tool for big data processing that provides a managed, scalable Hadoop **cluster**
  - Relies on S3 and HDFS
  - Used in a variety of applications, including ETL, clickstream analysis, real-time streaming, interactive analytics, machine learning

AWS Big Data Services

- The main component in EMR is a **cluster**
  - A collection of nodes (EC2s) and each node can be
    - Main
      - Responsible for managing the cluster
    - Core
      - Responsible for running tasks and storing data in the HDFS
    - Task
      - Optional node responsible only for running tasks

AWS Big Data Services

- Amazon EMR contains several layers:
  - Storage
    - Contains different file systems: HDFS, EMR File System, . . .
  - Cluster Resource Management
    - Responsible for managing cluster resources and scheduling jobs
    - Default is YARN
  - Data Processing Frameworks
    - Hadoop and Spark
  - Applications and Programs

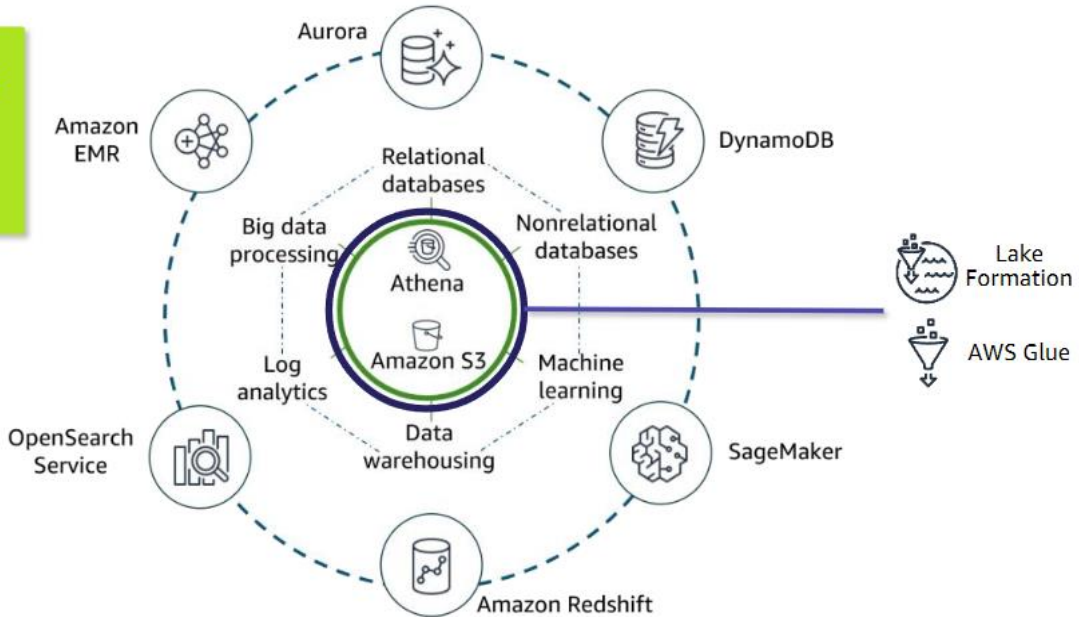
AWS Big Data Services

- AWS EMR clusters can be:
  - Persistent
    - For long running tasks
    - Default
  - Transient
    - Effective for periodic processing jobs
- EMR cluster can be launched:
  - Interactive mode
  - CLI mode
  - API mode

AWS Big Data Services

Key design considerations

- Seamless data movement
- Unified governance



Big Data Services on Clouds

- Big Data on Google Cloud Platform(GCP)
  - **Dataproc**
    - A fully-managed, highly scalable service
    - Runs Apache Hadoop, Apache Spark, Apache Flink, Presto, etc.
  - **BigQuery**
    - A fully-managed, serverless data warehouse that enables businesses to store and analyze massive amounts of data using SQL
  - **BigTable**
    - A distributed storage system developed by Google to store massive amounts of data and to scale up to thousands of storage servers
    - NoSQL database service

Big Data Services on Clouds

- **Dataflow**
  - A fully-managed cloud service that enables businesses to process and analyze streaming and batch data using Apache Beam.
- **Google Cloud Data Fusion**
  - A data integration service used to build and manage ETL (Extract, Transform, Load) data pipelines

Big Data Services on Clouds

- Microsoft Azure big data services
  - **Azure Synapse Analytics**
    - Provides a managed service for large-scale, cloud-based data warehousing
    - SQL support
  - **HDInsight**
    - Allows you to create clusters using Hadoop
    - Supports Interactive Hive, HBase, and Spark SQL, which can also be used to serve data for analysis.
  - **Azure Databricks**
    - Analytic service based on Apache Spark
  - **Azure Stream Analytics**
    - A serverless end-to-end streaming pipeline
    - SQL Support

- References
  - https://docs.aws.amazon.com/
  - Simplify Big Data Analytics with Amazon EMR. Sakti Mishra. 2022, Packt Publishing
  - https://aws.amazon.com/emr/