Big Data Processing Use Cases

Tessema Mengistu(PhD)

mengistu@cs.vt.edu

Outline

- Batch – ETL use case
- Realtime Stream use case
- IoT use case

Introduction

- Big data processing can be:
  - Batch Processing:
    - Processes data in batches, after it has been collected and stored in a data lake or warehouse
    - Example: genomic data processing, sentiment analysis
  - Stream processing:
    - Processes data as soon as it arrives, in real-time or near-real-time
    - High volume and velocity
    - Example: Fraud detection, IoT processing

Batch – ETL use case

- ETL - Extract - Transform - Load
  - Extract - data from different sources
  - Transform – into a format for processing or analysis
  - Load - to data in the correct state to the data warehouse or lake
- Batch ETL use cases are very common in big data processing

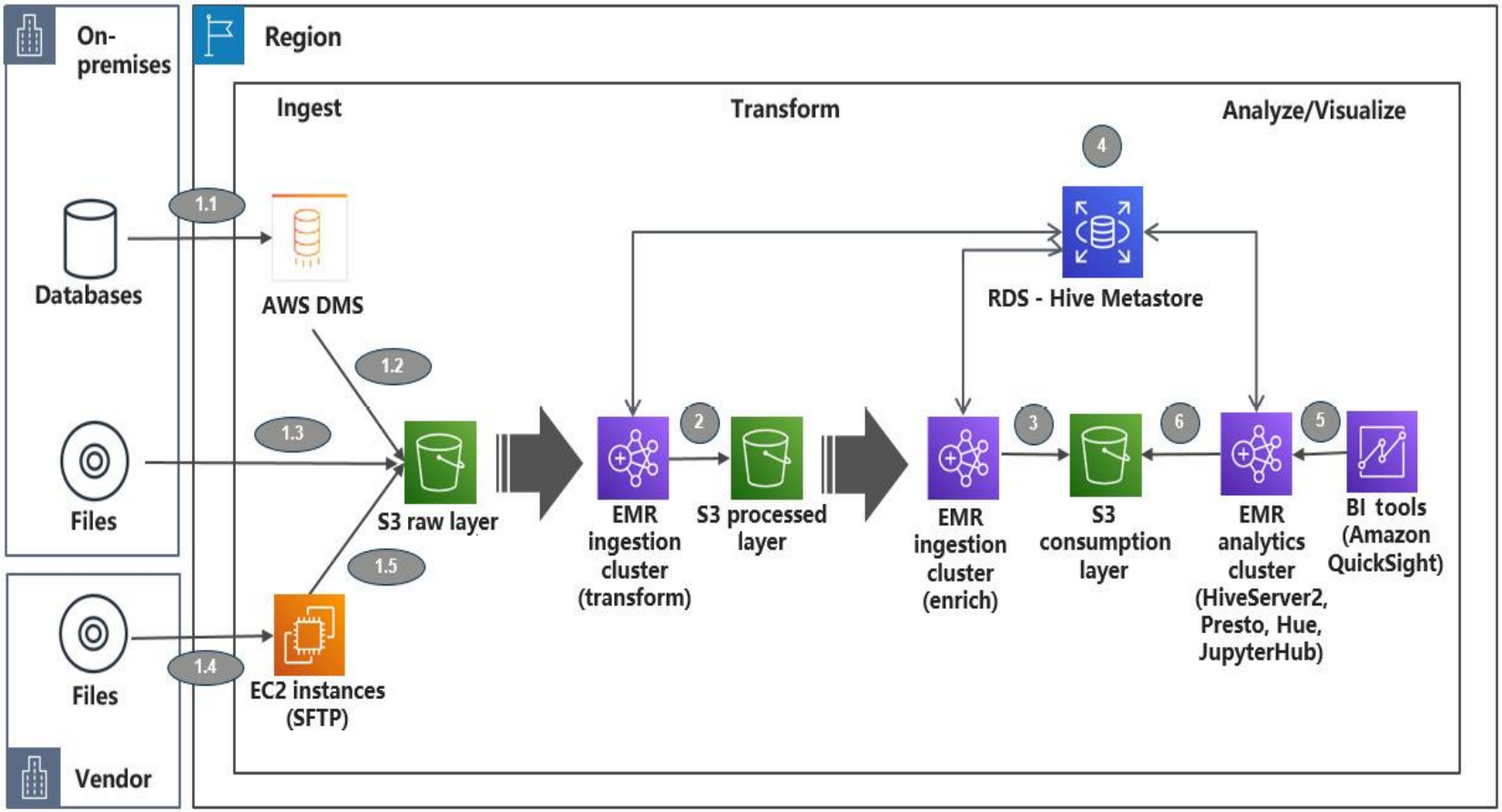
Batch – ETL use case

- Use case 1 - Batch ETL
  - Data sources
    - On-premises systems, which includes two data sources
      - A relational database
      - A filesystem
    - A vendor filesystem that uses SSH File Transfer Protocol (SFTP) to send files
  - Output Objective
    - Curating the data in an Amazon Simple Storage Service (S3) data lake, and then making the data available for consumption, where it should be able to accessible through SQL for analysis

Batch – ETL use case

- Steps:
  - Data migration from the source to AWS
    - Data Migration Service (DMS)
  - Clean and standardize the data - EMR
    - Multiple iterations
  - Processing - Hive/Spark
  - Analysis - BI tools

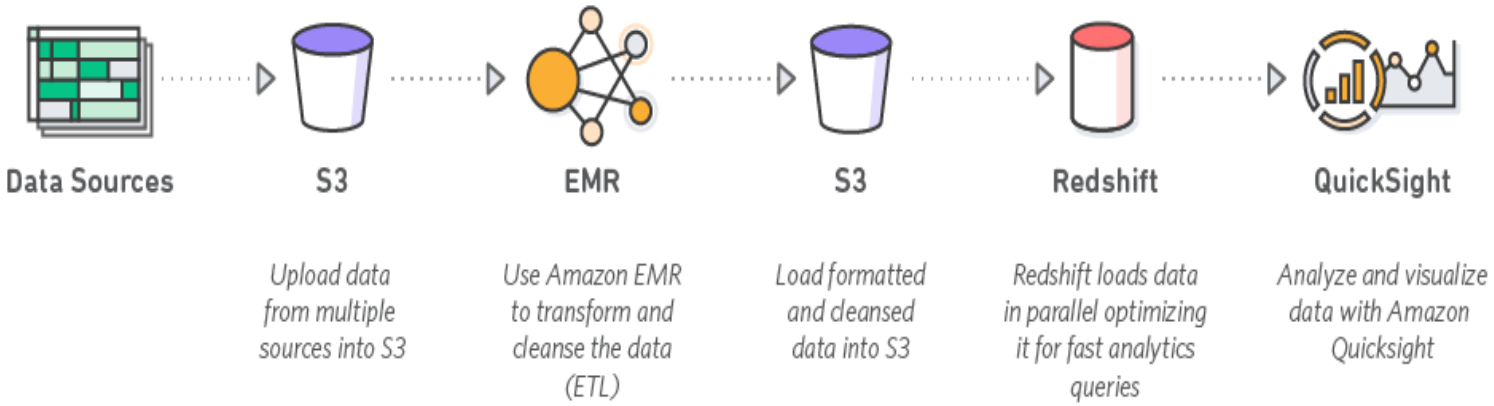
Batch – ETL use case



Source: Simplify Big Data Analytics with Amazon EMR

Batch – TEL use case

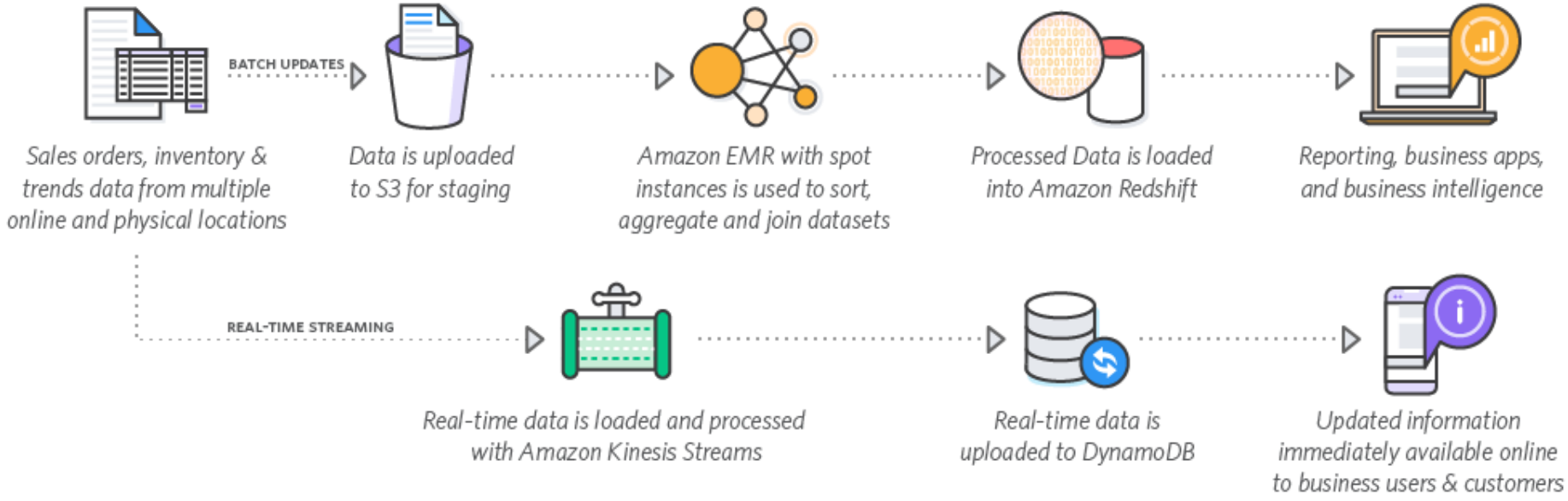
- Nasdaq
  - Amazon EMR uses Apache Hadoop framework to perform data transformations (ETL) and load the processed data into Amazon Redshift for analytic and business intelligence applications.



https://www.youtube.com/watch?v=O4wAH5FQjS8

Batch – ETL use case

- Redfin
  - Provides real estate listing & recommendations to millions of homebuyers
  - Redfin uses Amazon EMR with spot instances



https://aws.amazon.com/solutions/case-studies/redfin/

Batch – ETL use case

- Consideration
  - Keeping the raw data from the source system persistently
- Best practices
  - Data Partitioning
  - Transient EMR
    - Reduce cost

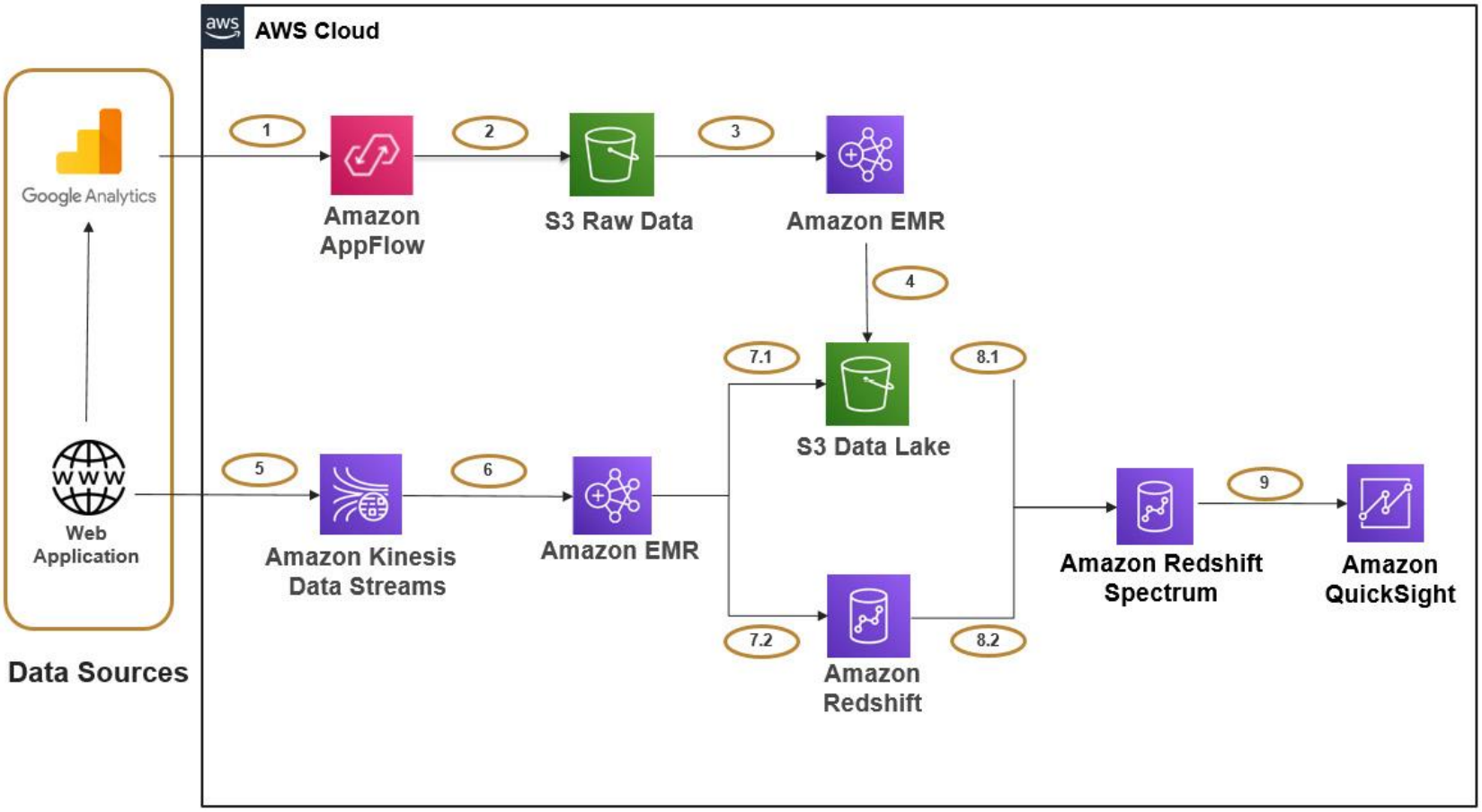
Realtime Stream use case

- Use case 2 - Real-time streaming clickstream application
  - Data source
    - A multinational retail store website with a huge userbase and traffic
    - The website uses Google Analytics
  - Objective
    - User session-based analytics
  - You need to stream user clicks and Google Analytics data and do the analyses

Realtime Stream use case

- Possible Steps:
  - Data collection
    - From Google Analytics
      - Amazon AppFlow
    - User clicks
      - Amazon Kenesis
  - Clean and standardize the data - EMR - Spark
    - Multiple iterations
  - Processing - Redshift
  - Analysis - BI tools

Realtime Stream use case



Source: Simplify Big Data Analytics with Amazon EMR

Realtime Stream use case

- Hearst Corporation
  - Monitors trending content for over 250 digital properties worldwide
    - 30TB of data per day
  - Using an architecture that includes Amazon Kinesis and Spark running on Amazon EMR, Hearst corporation delivers real-time insights

Send clickstream data to Kinesis Streams

Kinesis Streams stores and exposes clickstream data for processing

Custom application built on Kinesis Client Library makes real-time content recommendations

Readers see personalized content suggestions

https://aws.amazon.com/solutions/case-studies/hearst/

Realtime Stream use case

- Consideration
  - How long to keep the raw data
- Best Practices:
  - Scalability
  - Fault tolerance

Realtime Stream use case

- Use case 3 – Log analytics
  - Data source
    - Server logs
      - EC2 servers generating logs that include CPU, memory usage, error logs, or access logs
    - Application logs
      - Each application is generating debug or error logs
      - For example, Java applications are generating logs through the Log4j framework
    - Apache and NGINX logs
      - When applications are deployed or accessed through Apache or NGINX servers, they also generate access logs or error logs

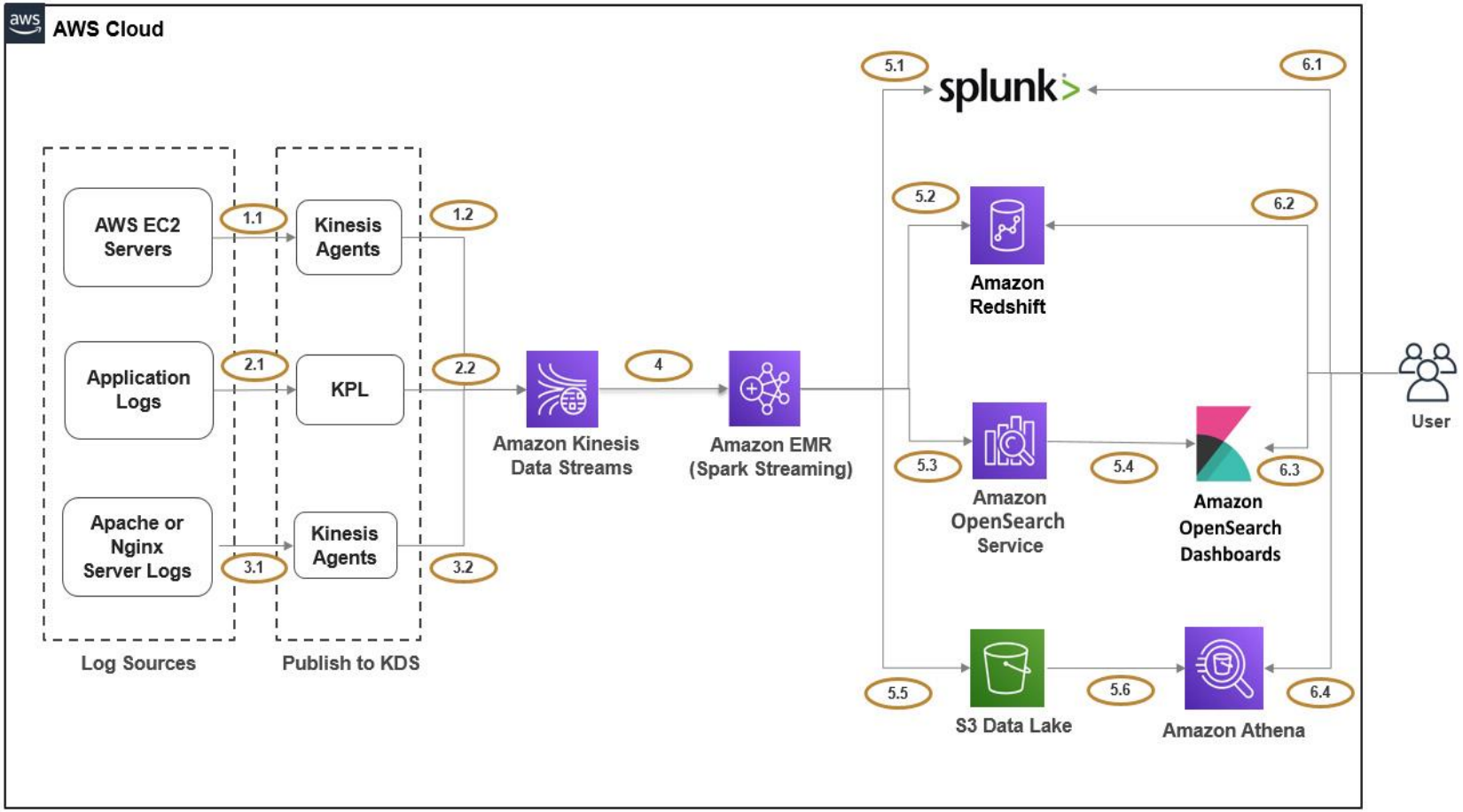
Realtime Stream use case

- Objective
  - No unauthorized access
  - Find common failure pattern
- You need to stream logs
  - Different schema

Realtime Stream use case

- Possible Steps:
  - Data collection
    - Form EC2 servers, application logs
      - Amazon KDS
  - Clean and standardize the data - EMR – Spark streaming
    - Different schemas
  - Processing - Redshift
  - Analysis - search tools

Realtime Stream use case



Source: Simplify Big Data Analytics with Amazon EMR

Realtime Stream use case

- Best Practices
  - Scalability

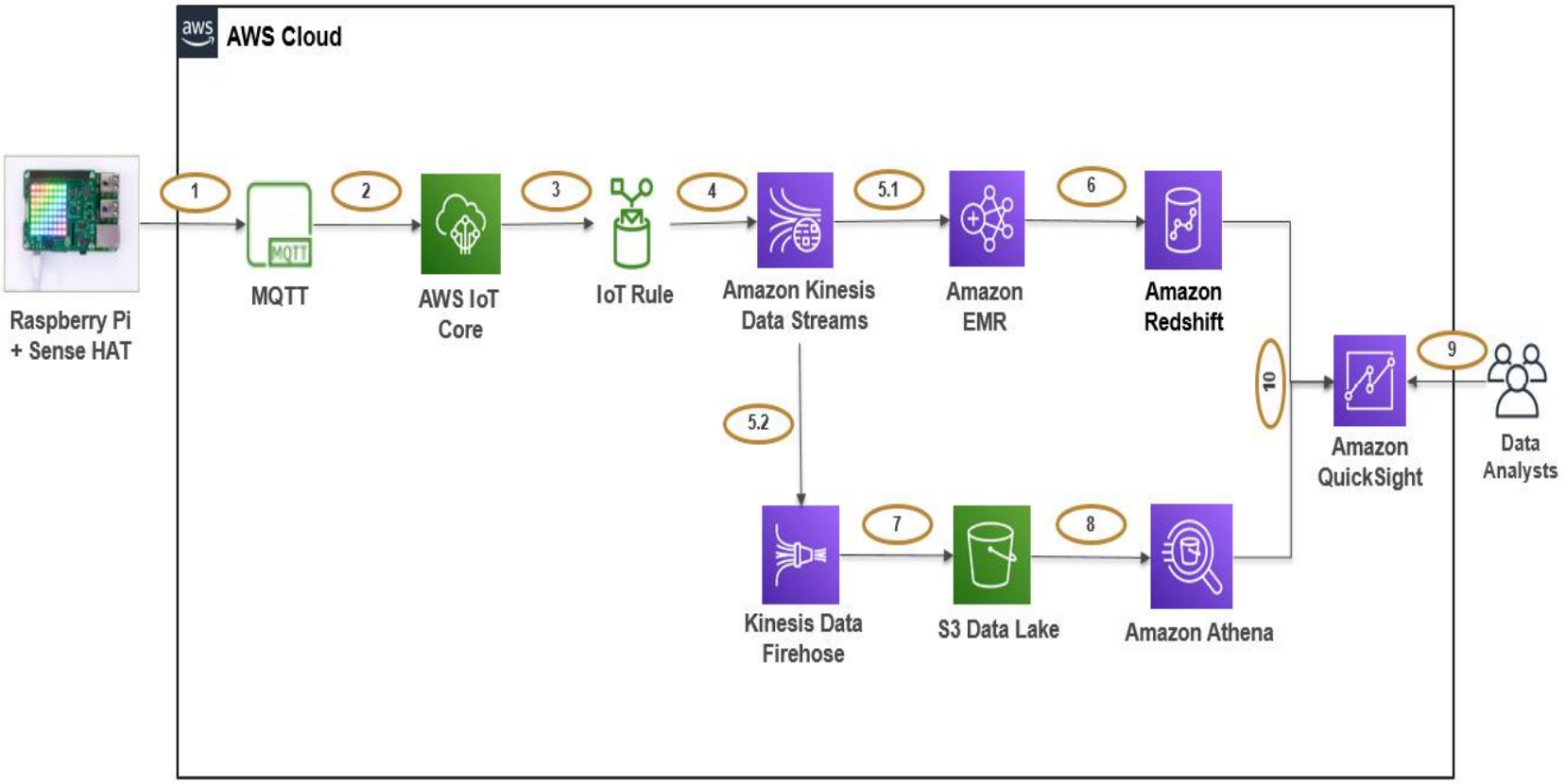
IoT use case

- Use case 4 - IoT
  - Data source
    - IoT devices to track electric usage at homes or offices
    - Real time
  - Objective
    - Derive insights, and then provide analytical reports and recommendations to their users

IoT use case

- Possible Steps:
  - Data collection
    - IoT devices send data directly
      - KDS
  - Spark-streaming for processing
  - Processing - Amazon Athena
  - Analysis - BI tools

IoT use case



Source: Simplify Big Data Analytics with Amazon EMR

IoT use case

- Best practices
  - Buffering data

- References
  - Simplify Big Data Analytics with Amazon EMR. Sakti Mishra. 2022, Packt Publishing
    - Chapter 3