# Scenario

Your company serves **media content** to affiliate websites and mobile applications. How would you build a system to help your company's analysts optimize served content?

The solution is fairly simple:

From Web and mobile client -> Kinesis Data Firehose (near real time) -> Raw Data bucket (S3) -> Amazon EMR with Apache Spark (for ETL job) -> Intermediate data bucket -> RedShift

You need to choose between real time / Micro-batches (appropriate for here!) / Batch

## How to collect data?

Use Kinesis Data Firehose to Transform data! (delivers stream data into Amazon S3 bucket)

You might remember the configuration with routing to backup S3 bucket in case of transformation failure (optional).

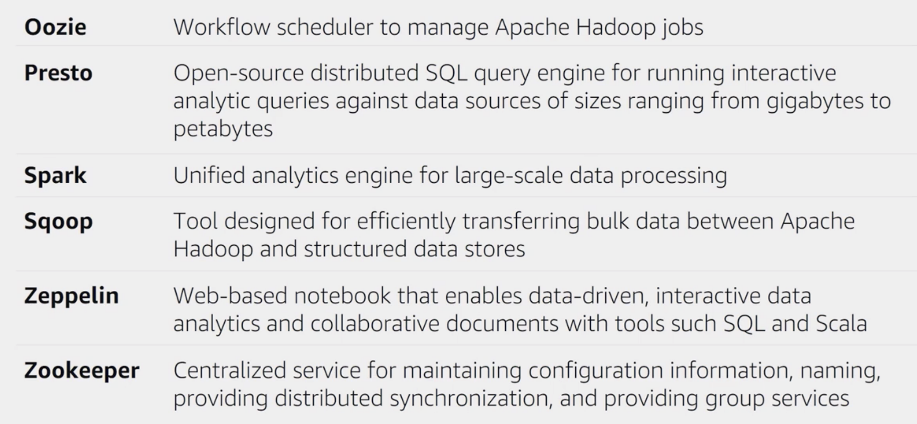
Transformation here refers to compression, encryption with KMS, convert using Lambda. Keep in mind: **Kinesis firehose can transform one record at a time.** So for more complex transformation, the solution is EMR with Spark, not Lambda.

## ETL

Source Data -> Amazon EMR running Apache Spark -> Destination Data

### There are stack of libraries included in Spark

1. Spark SQL : query structured data inside Spark program (it can read data from existing Hive installation)
2. Spark Streaming : let you re-use the same code for both streaming data and batch processing! You can join streams with historical data and make ad-hoc queries on stream state (ingest data from Kinesis, process it, push data to file systems, databases, dashboards. . . )
3. MLLib : Machine Learning library
4. GraphX : graph analysis library



| | |
|---|---|
| **Oozie** | Workflow scheduler to manage Apache Hadoop jobs |
| **Presto** | Open-source distributed SQL query engine for running interactive analytic queries against data sources of sizes ranging from gigabytes to petabytes |
| **Spark** | Unified analytics engine for large-scale data processing |
| **Sqoop** | Tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores |
| **Zeppelin** | Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with tools such SQL and Scala |
| **Zookeeper** | Centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services |

Figure 1: image

**There are also Sandbox Applications**

**Cost optimization**

In terms of lifetime: Run EMR cluster only when it is needed! Permanent vs Transient clusters

In terms of instance cost: If you need your EMR cluster running 24/7, Use spot instance for task instance.

## RedShift

MPP (Massively Parallel Processing)

Each cluster has one leader node, and two or more compute nodes

Each compute nodes has a number of slices. Each slice is an independent partition of data, and they work in parallel.

Some queries are exclusive to Leader node, and the others are processed in compute node.

For key distribution, small dimension table doesn't benefit from ALL distribution.

Redshift store data in blocks, and each block stores metadata of minimum and maximum of the sort key

### Class Discussion

When should you use Amazon EMR? -> if you use custom code to process and analyze large data sets with big data processing frameworks like HBase, Spark, . . . Gives full control of configuration of cluster and the software you wanna install.

When should you use Amazon Redshift -> Perform complex queries on massive collections of structured data. Data warehouse solution is designed to pull from multiple data sources for analysis. To ensure the reporting is consistently accurate across the entire company, data warehouse stores data in highly structured fashion.

### Class Discussion 2

What are the functions of the Redshift leader node?

A: It acts as a SQL endpoint, stores metadata, and generate and coordinates query execution plan.

### Class Discussion 3

What are the AWS services that can be used to load data into Amazon Redshift

DynamoDB, EMR, Data pipeline, Firehose, S3, ..

## Sample exam question

Your manager has asked you to upload some data to S3. She asked you to make sure that the data is encrypted at rest and can be easily loaded into Amazon EMR for analysis. How should you protect your data?

A: Upload your data to Amazon S3 and choose an AWS KMS key to encrypt the data. Give your Amazon EMR cluster an IAM role that grants access to the data and the AWS KMS key.