Analytics with DyanmoDB

Ivan Mushketyk

@mushketyk brewing.codes

Overview

Impossible to run analytics query with DynamoDB

Can be interested in questions like:

- What is the most often bought item
- Who is the most active customer

Will implement similar queries using:

- Redshift
- EMR & Apache Hive

Using AWS Redshift

What is Redshift

How it works

How to use it with DynamoDB

What is Redshift

Data warehouse solution

PostgreSQL dialect

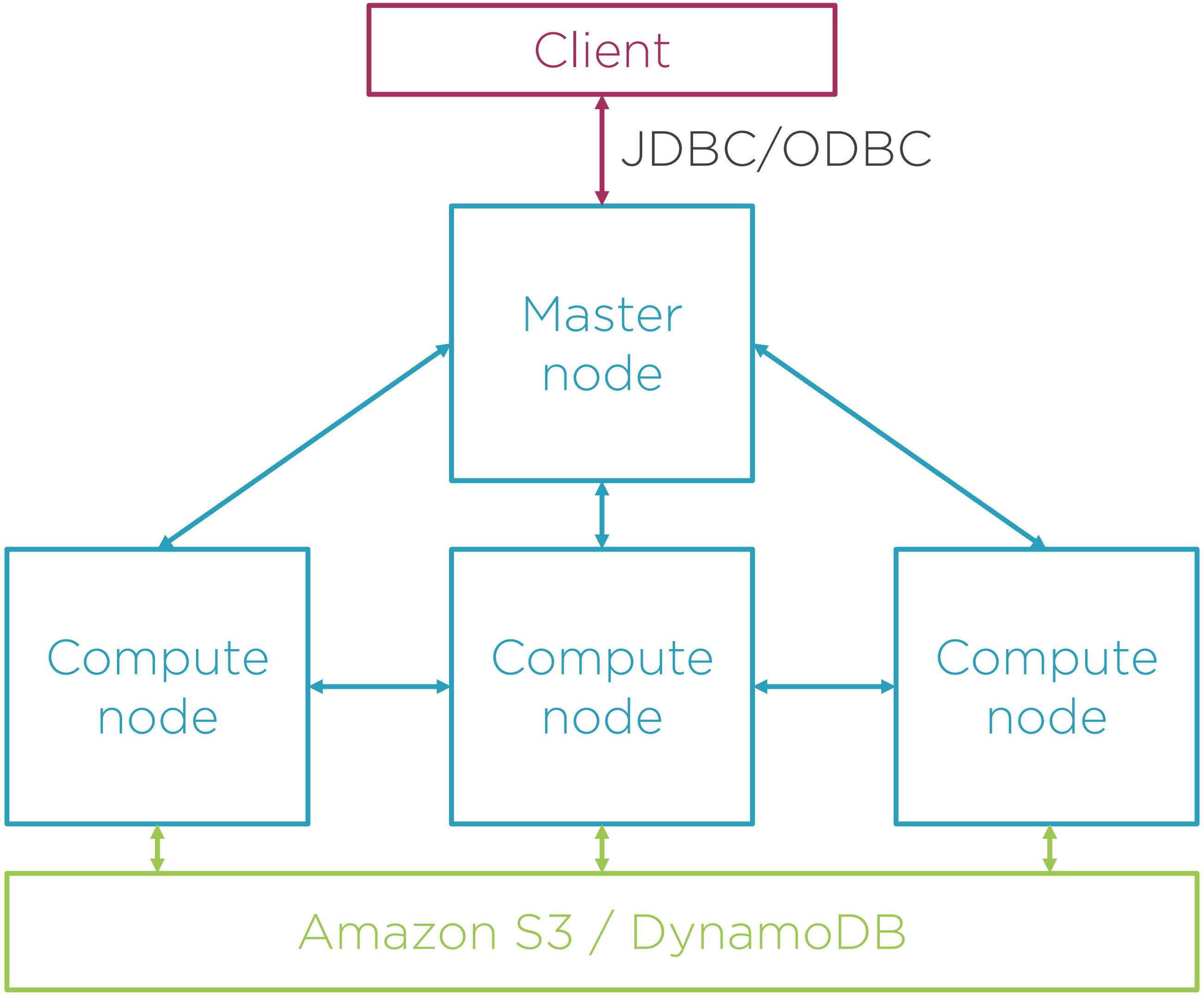
Fully-managed, scalable

Petabyte scale

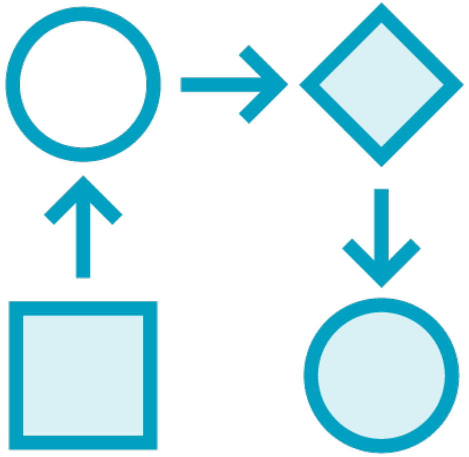
Allows to perform complex queries on DynamoDB data

Works on copy of data

How Redshift Works



Redshift Workflow



Launch Redshift cluster

Copy DynamoDB data

Perform SQL queries

Copy Data from DynamoDB

```
copy messages
from 'dynamodb://Comments'
credentials 'aws_access_key_id=<Access-Key>; aws_secret_access_key=<Secret-Access-Key>’
readratio 50;
-- Alternativelly
iam_role 'arn:aws:iam::0123456789012:role/AwsRole';
```

Redshift Gotchas

Tables:

- Limited to 127 characters
- Can’t contain ’.’ or ‘-’]
- Table can’t be named with reserved word

DynamoDB types do not correspond directly to Redshift types

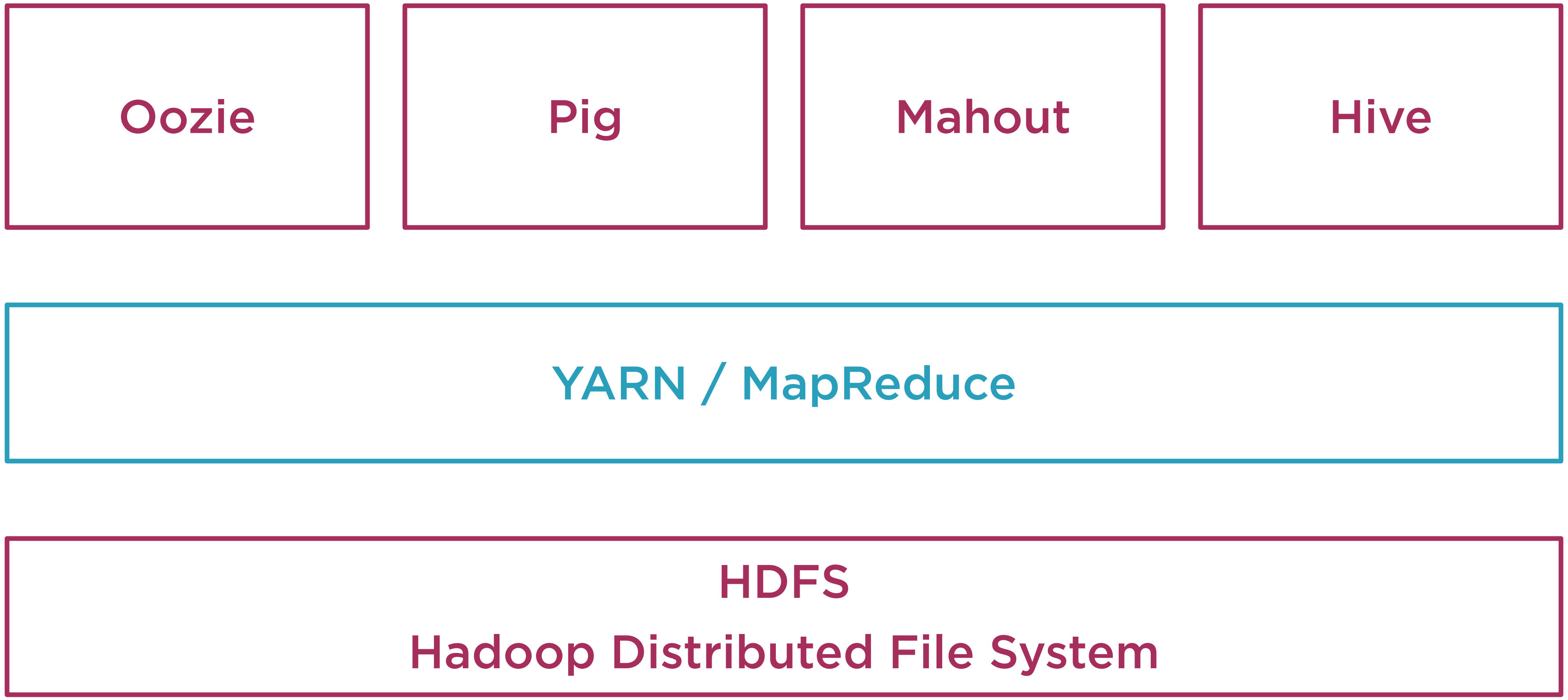
How to Use EMR Hive

What is Apache Hive

What is EMR

How to query DynamoDB data

Apache Hadoop Stack



Oozie

Pig

Mahout

Hive

YARN / MapReduce

HDFS

Hadoop Distributed File System

What Is Apache Hive?

Tool for ad hoc analysis of distributed data

Support SQL-92 specification

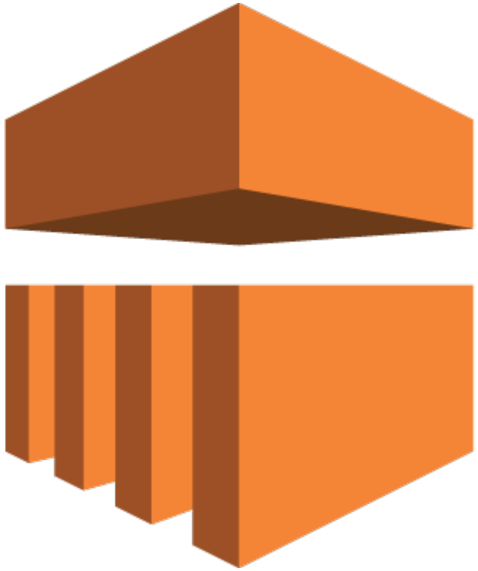
Converts SQL into MapReduce jobs

Not a database

High latencies

Can read from DynamoDB

What Is Amazon EMR?



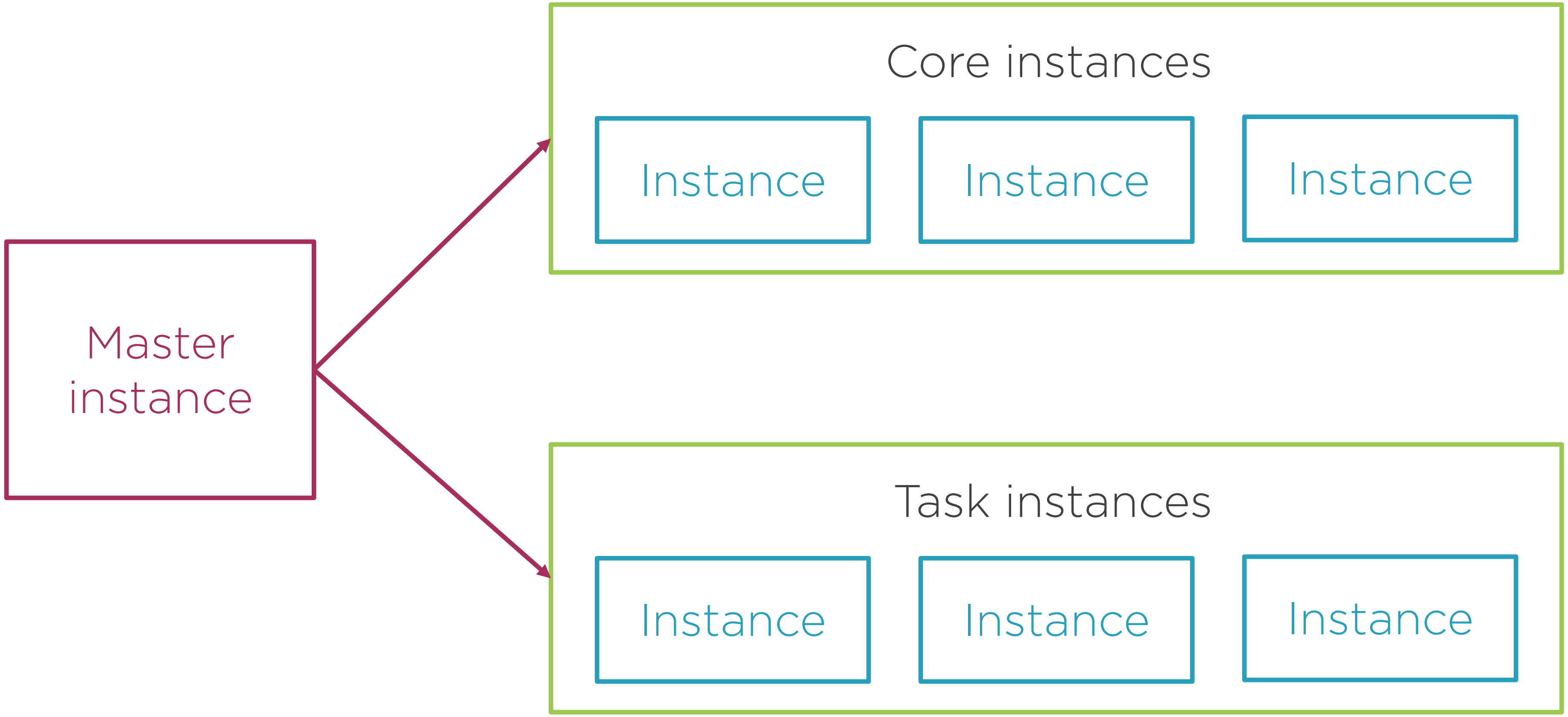
Service for running Hadoop projects

Elastic service

Supports multiple tools

- Spark
- Flink

How Does It Work?



Where to Query Data From?

Directly from DynamoDB

Import DynamoDB data to S3

Import DynamoDB data to HDFS

How to Query Data with Hive?

Start EMR cluster

Connect to master node

Create external table

Write query

Create External Table

```
CREATE EXTERNAL TABLE ddb_messages
    (item_id  STRING,
    message_id STRING,
    message STRING,
    timestamp BIGINT)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler’
TBLPROPERTIES(
"dynamodb.table.name"="Messages",
"dynamodb.column.mapping"="item_id:ItemId,message_id:MsgId,message:Msg”
);
```

Pros and Cons of Querying DynamoDB

- + Queries latest data
- + Can use JOIN, GROUP BY, aggregation functions, etc.
- − Consumes table’s RCUs
- − May limit number of queries

Copy DynamoDB Data to S3

```
INSERT OVERWRITE DIRECTORY 's3://aws-logs-123456789012-us-west-2/dynamo-copy' SELECT * FROM ddb_messages;

CREATE EXTERNAL TABLE ddb_messages
   (item_id  STRING,
   message_id STRING,
   message STRING,
   timestamp BIGINT)

LOCATION 's3://aws-logs-123456789012-us-west-2/dynamo-copy';
```

Summary

- Data analytics with DynamoDB
- DynamoDB does not support complex queries
- Can copy data to Redshift
- Can use EMR Hive
- Can copy data to S3/HDFS and use Hive