DATA+AI SUMMIT 2022

Data Processing using Spark on Cloud

A Low cost Data Management Ecosystem with Apache Spark at Core

ORGANIZED BY databricks

Shariff Mohammed
Distinguished Engineer, Capital One

Session Context

Agenda

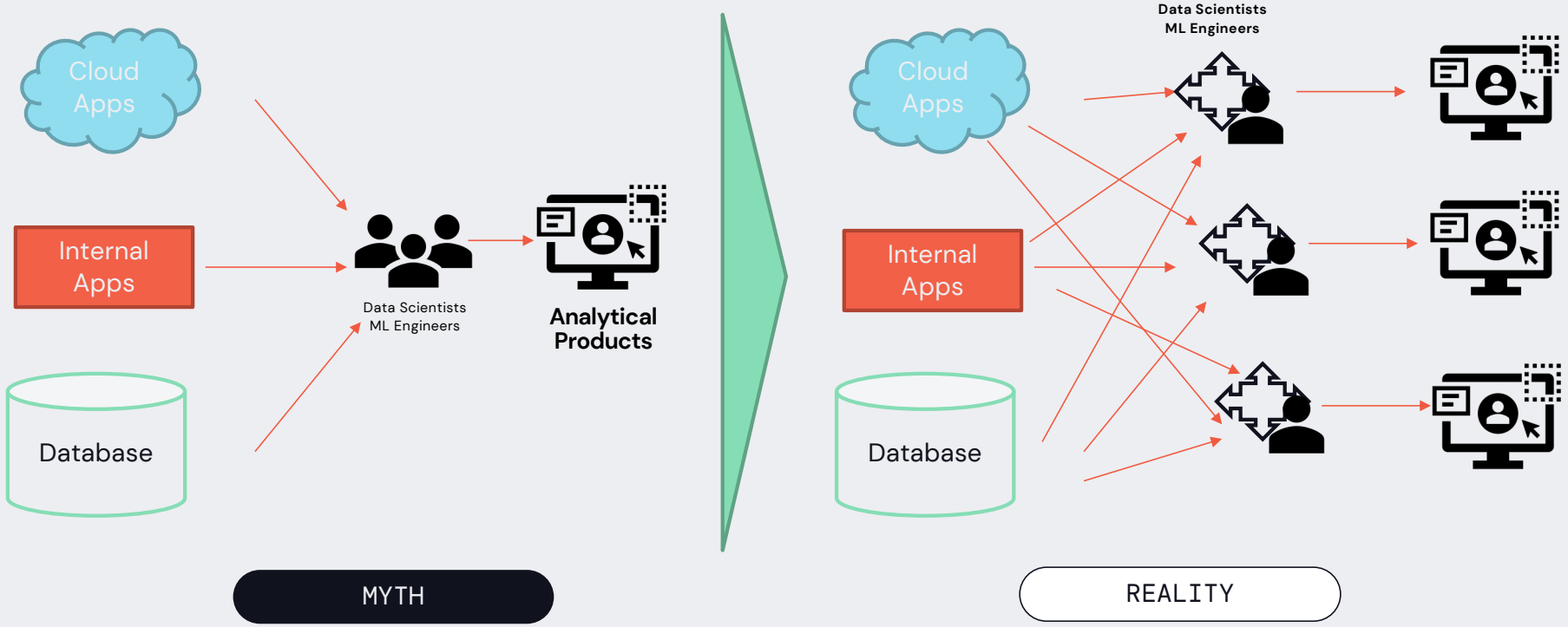
- Role of Data Processing in the Data Management Ecosystem.
- Evolution of ETL Tools (On-Prem to Cloud-era)
- Data Processing (ETL) Architecture on Cloud
- How Spark powered our Ecosystem
- The gist of the story

Data Processing

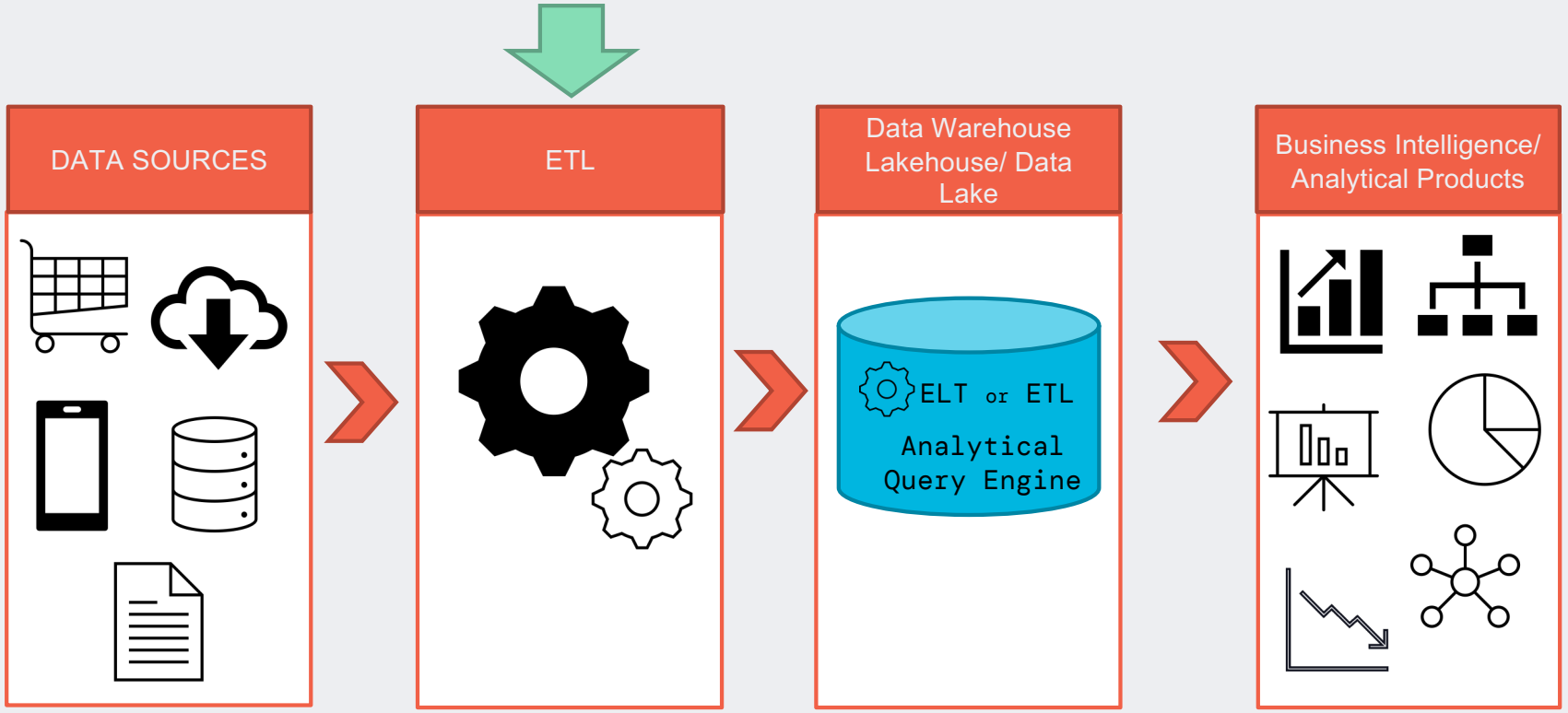
Where does it fit in the data ecosystem?

Myth buster

A popular belief is that source data is ready for usage



Data Preparation before usage can drive efficiencies in many ways



Evolution of ETL Tools

Evolution of ETL Tools

Legacy (Pre-Cloud Era)

- Enterprise-Scale tools were License based.
- Cost associated was with per Developer License & additional license for Server cost
- Niche knowledge to code (Developers were Tool-specific experts)

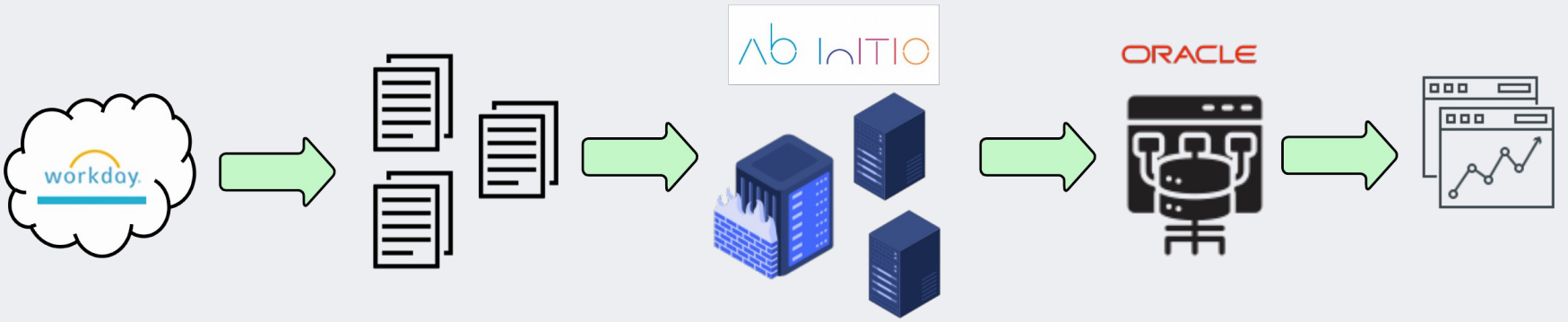
Current (Cloud)

- Both Open-Sourced and Enterprise License Tools available now.
- Open Source License based tools are free to use and can scale up to Enterprise usage as well.
- With easy to learn OS based languages like Spark, more talent available in the industry

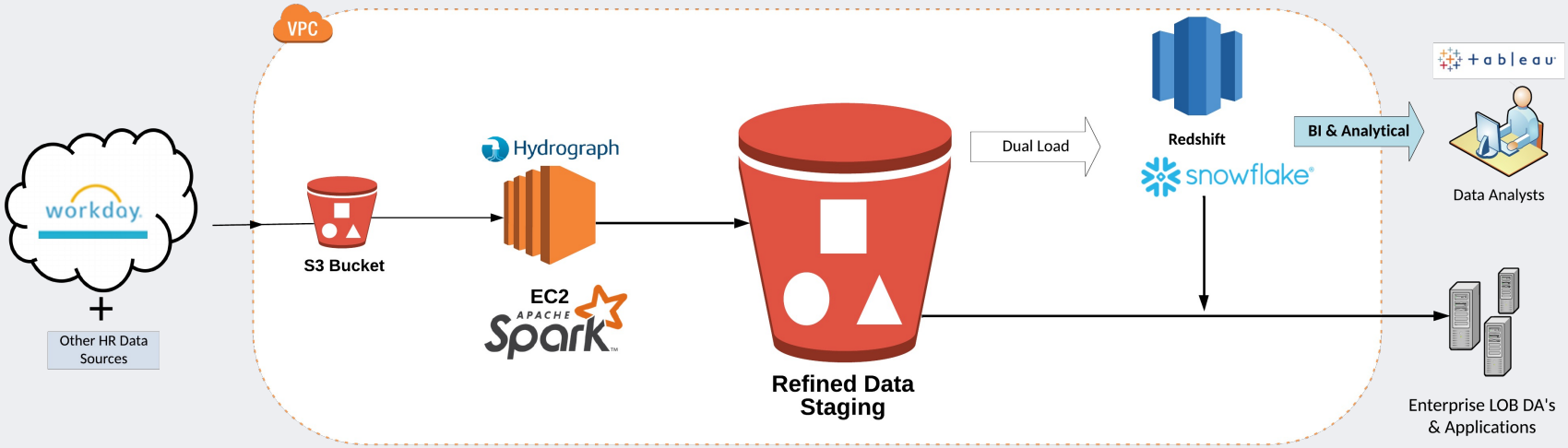
Building Analytical Platform on Cloud

Legacy EcoSystem –On Prem

Analytical Platform built to drive people decisions for the company



Migration to Cloud Current Data Ecosystem



Infrastructure Management

Infrastructure Automation

HashiCorp Terraform

AWS CloudFormation

DEV OPS

DevOps

CI/CD Tools

Jenkins

MEND

sonarqube

Monitoring

Job Scheduling

AROW

Infrastructure Monitoring

DATADOG

New Relic.

Log Monitoring

Incident Monitoring

Data Storage & Governance

Amazon Redshift

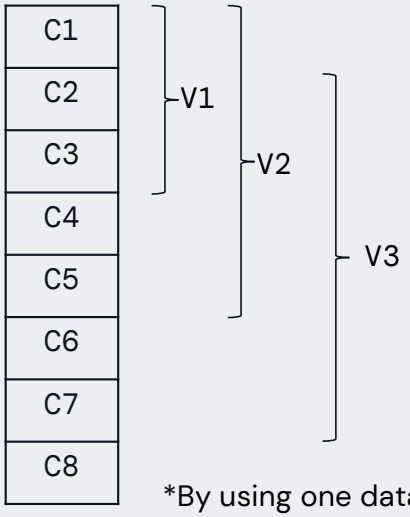
snowflake

OKERA

DATA+AI SUMMIT 2022

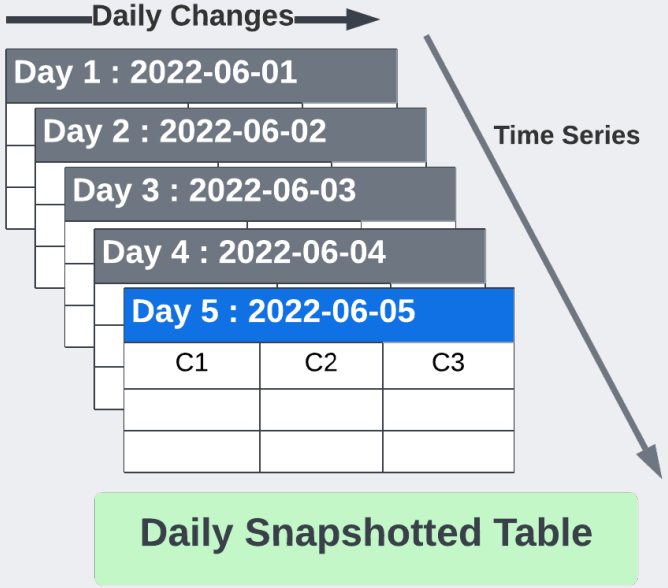
Reaping the benefits of Cloud Migration

Data Protection – Improved FGAC controls



*By using one dataset, we can share different data elements using Hive based solutions without creating redundant copies of data.

More Processing Power – History of Histories



Daily Snapshotted Table

Benefits using Spark

How Spark powered our Ecosystem

- Ease of Coding
- Infrastructure Agnostic
- Multi-threaded Processing
- Open Source License -> Cheaper cost

Ease of Coding

As Native Spark

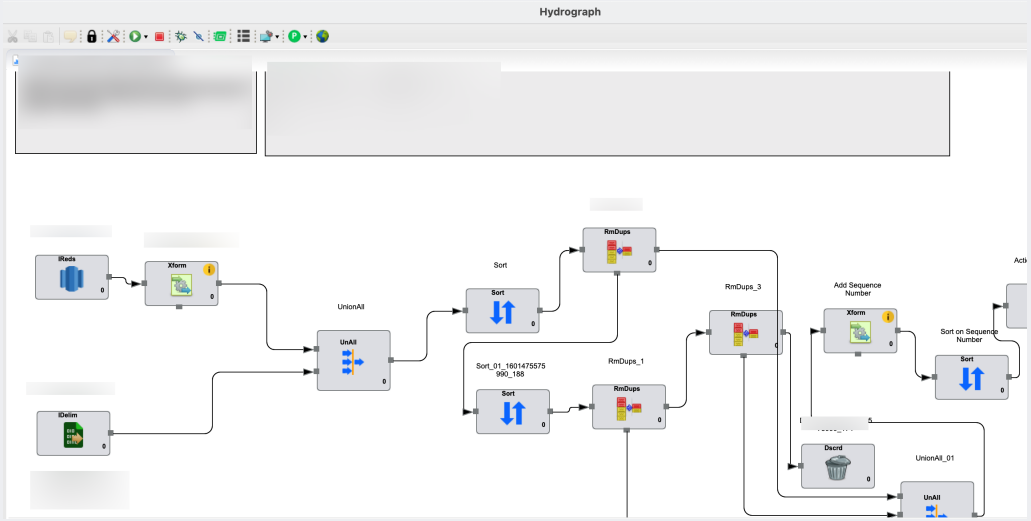
```
// Open connection to SQL Server database
SQLServerConnection Conn;
Conn = new SQLServerConnection("host=nc-star;port=4100;User ID=test01;
Password=test01;Database Name=Test");
try
{
Conn.Open();
Console.WriteLine ("Connection successful!");
}

```

```
./bin/spark-submit \
    --deploy-mode cluster --master yarn \
    --class org.apache.spark.examples.SparkPi \
    /spark-home/jobs/jars/jobname_versionxx.jar
```

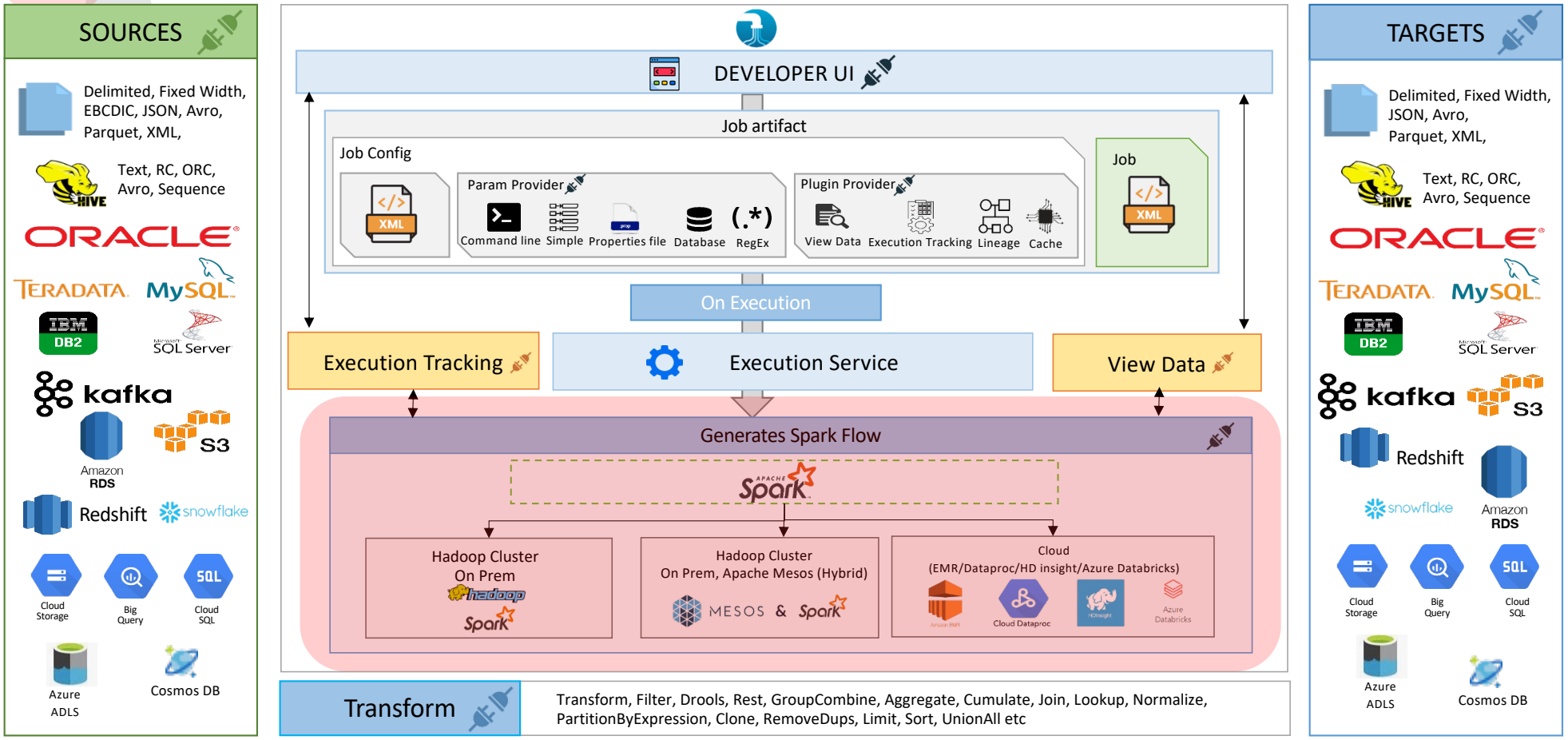
Ease of Coding

As Hydrograph



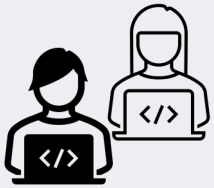

```
./bin/spark-submit \
    --deploy-mode cluster --master yarn \
    --class org.apache.spark.examples.SparkPi \
    /spark-home/jobs/jars/jobname_versionxx.jar
```

Hydrograph Pluggable Architecture

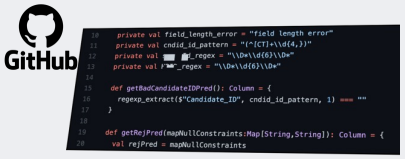


Versatile Infrastructure with Apache Spark

Development



Developers



Spark Code

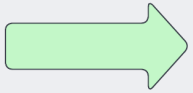
{..CI/CD..}

Executable
Spark Code

Deployment

Code Artifacts

Software



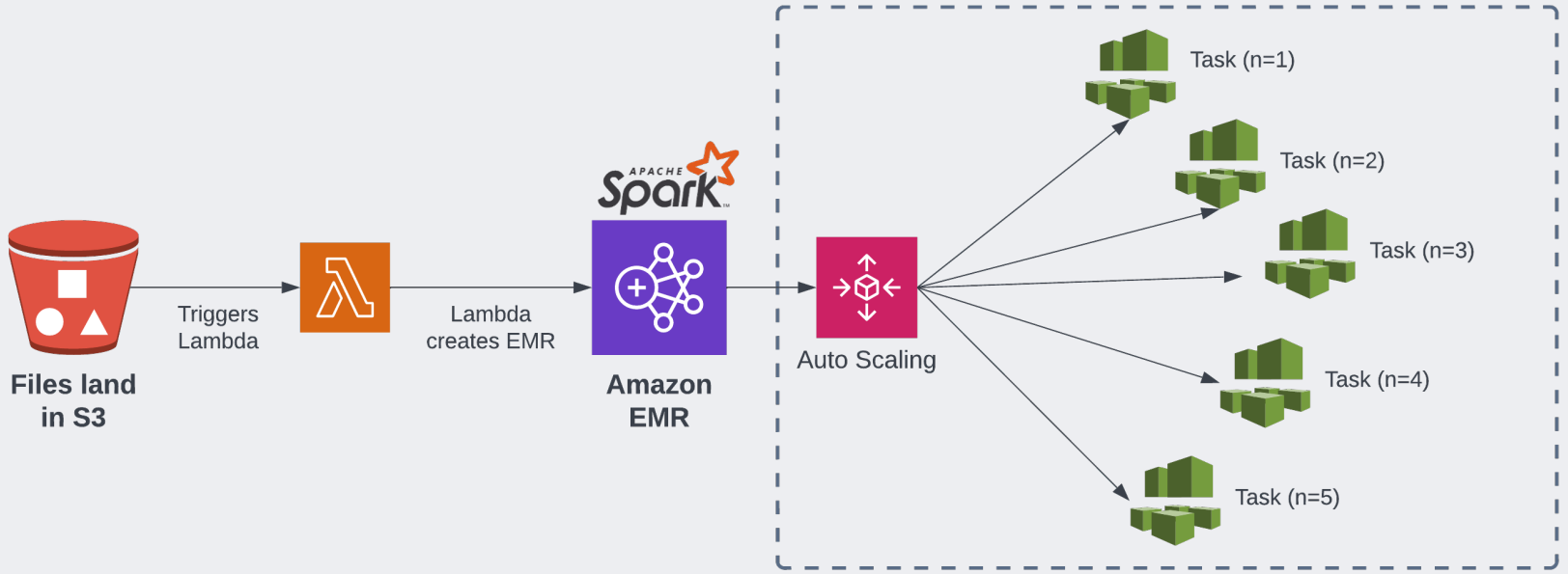
Amazon EC2

Amazon EMR

Amazon ECS

Scalability with Spark

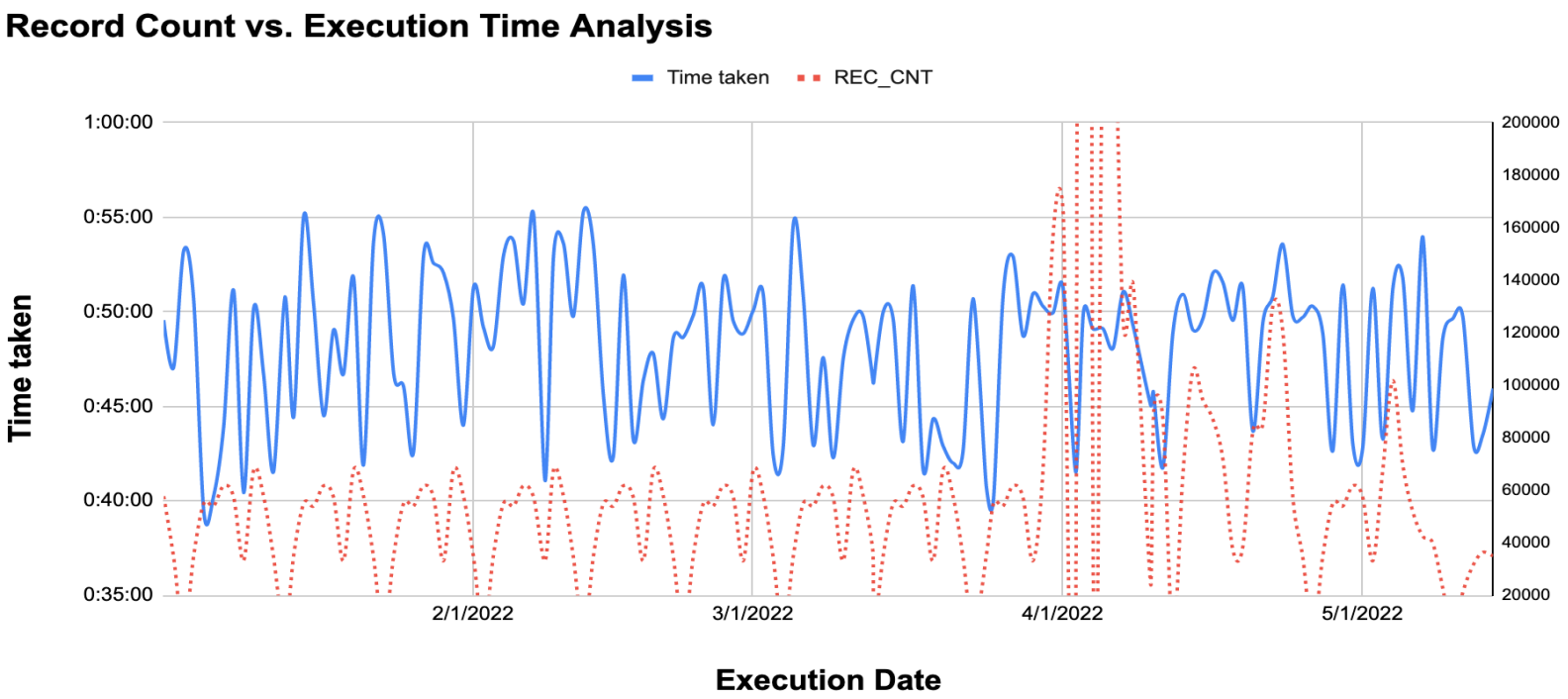
Multi-thread Processing on Spark EMR Cluster



Scaling up to multiple nodes will not require to be programmed, instead managed automatically managed on EMR Spark Cluster

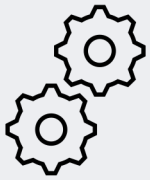
Scalability with Spark

Execution time is almost constant even when input record volume increased by 30x



Cost of ETL (Data Processing)

120+
Raw Files



700+
ETL Jobs

Amazon EC2 Instance

ec2 type: r5.8xlarge

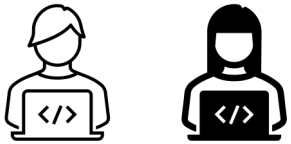
~$1000
per month

Gist of the story

Summary

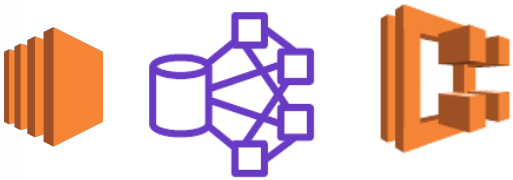
Benefits of building Spark-based ecosystems

Ease in Talent Hiring



Data Engineers are not ETL tool specific anymore

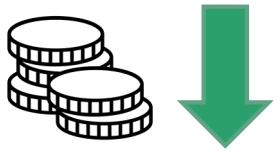
Infrastructure Agnostic



EC2 EMR ECS

Code portable to any popular data processing Infrastructure

Lower Costs



~$1000/month or

~$30/day

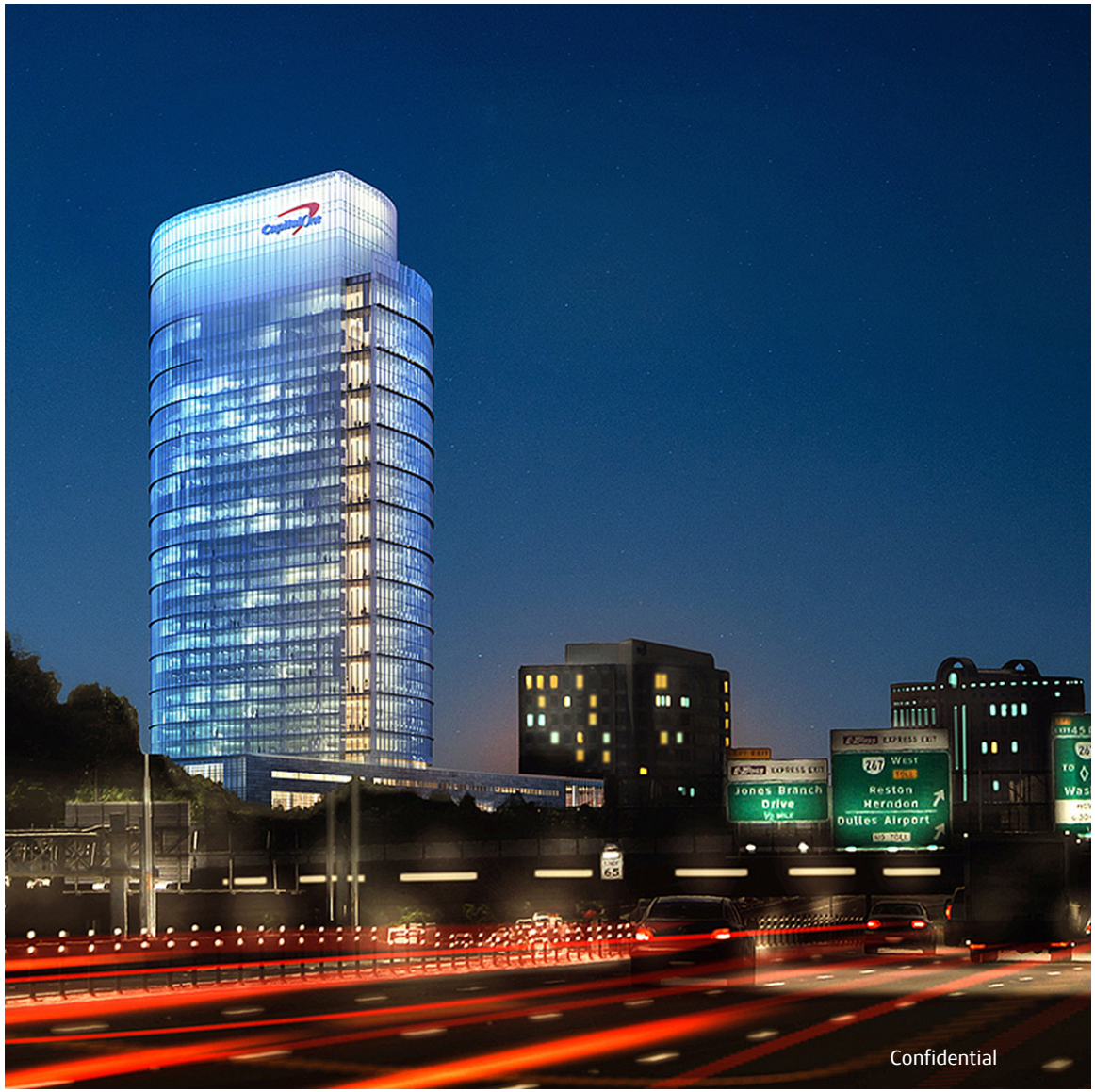
~700 jobs

"Our company was founded on a simple principle; recruit great people and give them the opportunity to be great."

–Rich Fairbank

WE ARE HIRING:

https://www.capitalonecareers.com/



DATA+AI
SUMMIT 2022

Thank you

Shariff Mohammed

Distinguished Data Engineer

Email: shariff.mohammed@capitalone.com

LinkedIN: https://www.linkedin.com/in/shariff-mohammed-96761328/