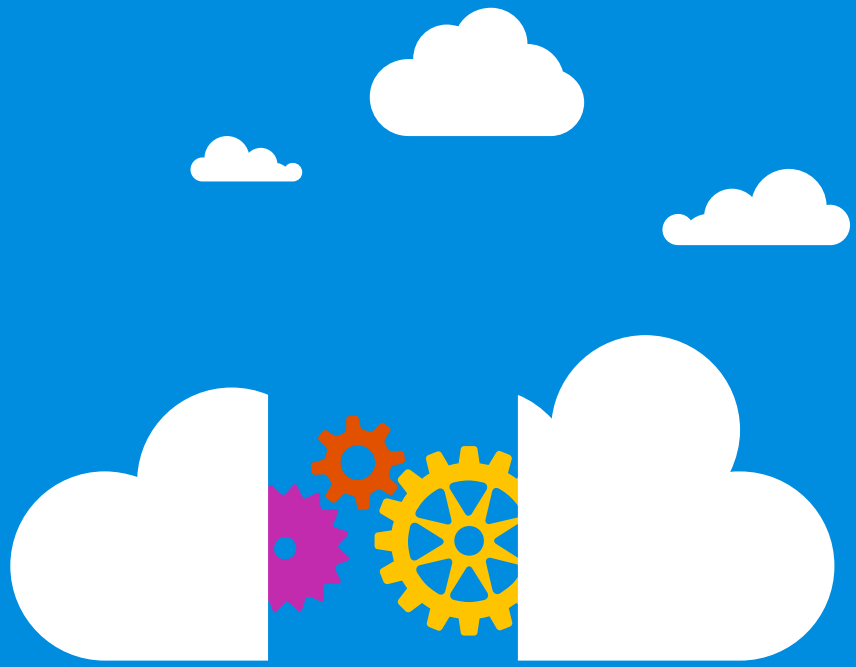
Solving low latency query over big data with Spark SQL

Julien Pierre



Microsoft

Agenda

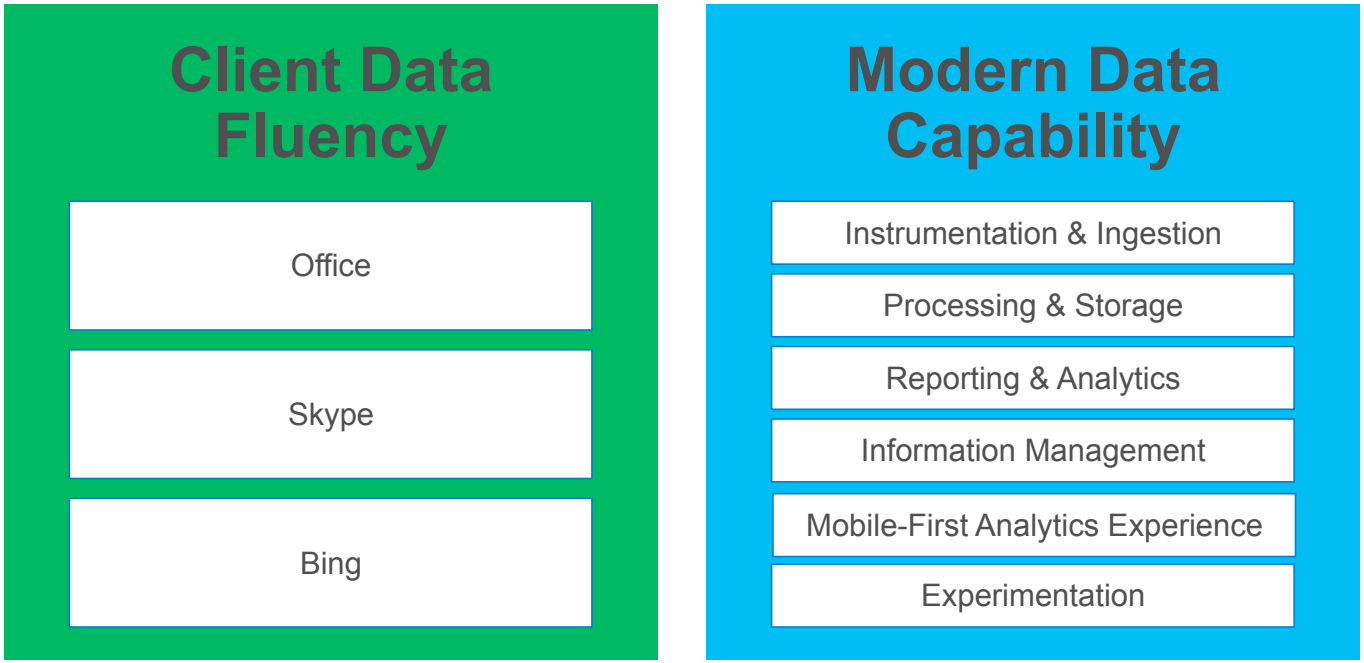
- Intro
- Project’s mission
- Use case
- Learnings
- Implementation details
- The future of Spark in ASG

Who am I?

Julien Pierre
朱力安

- Senior Product Manager
- ASG Shared Data team
- Beijing
- Working on:
  - Spark
  - Driving Microsoft internal Spark community
- Worked on:
  - Office 365 Admin
  - Office 365 Customer Facing Reporting
  - Open sourced the Office 365 Splunk App

What we do


Client Data Fluency
Office
Skype
Bing
Modern Data Capability
Instrumentation & Ingestion
Processing & Storage
Reporting & Analytics
Information Management
Mobile-First Analytics Experience
Experimentation

Project mission

Deliver interactive analytics capabilities for ASG teams to build great products

Increase analyst and engineer productivity

Capitalize on our data

Leverage open technology to run a stable service

Use Case

Where does Spark SQL help us?

Data issue troubleshooting use case

Kuntao Yu, engineer

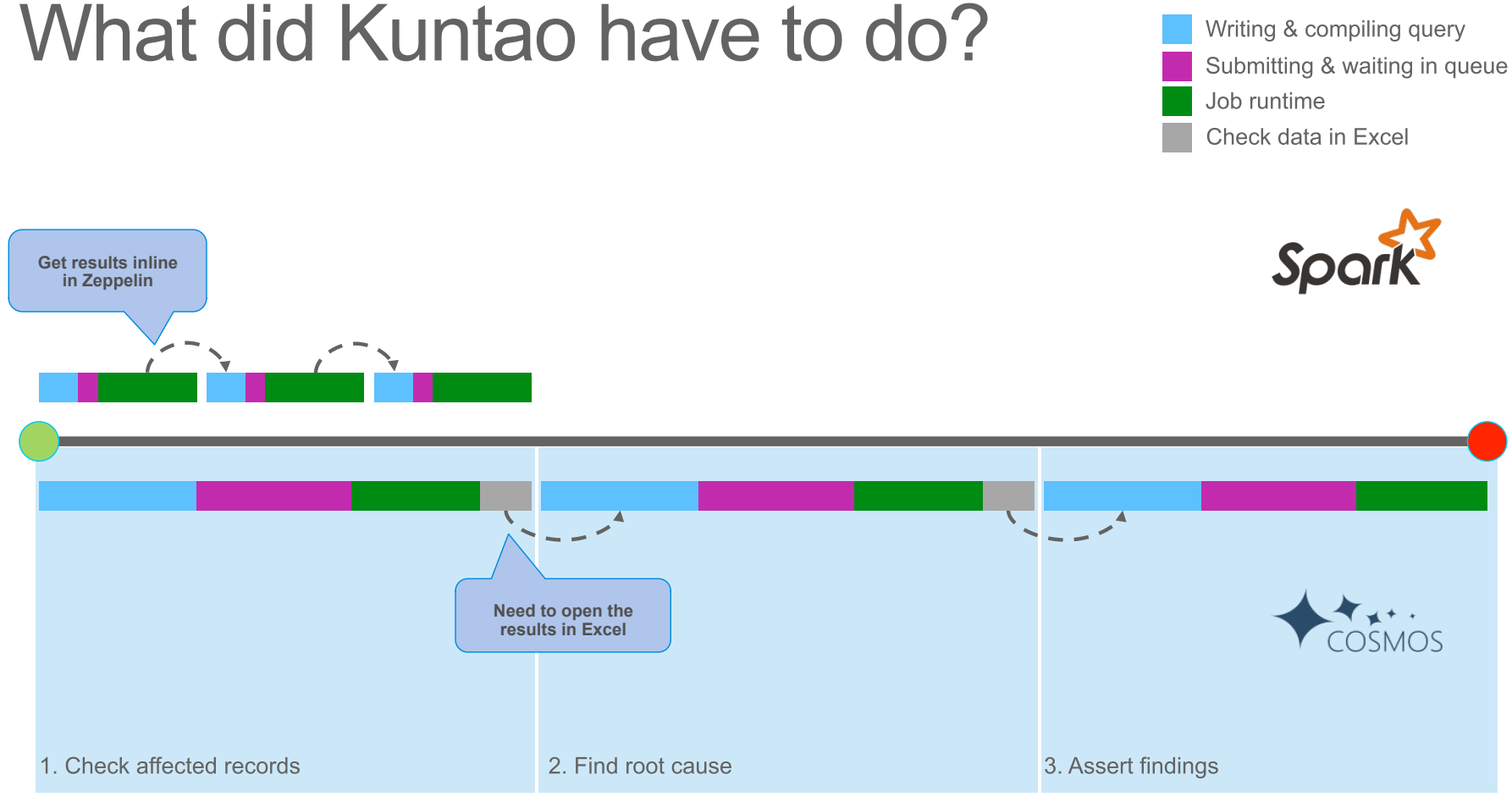
Customer found an issue with the data

"*Why is the Created Date field = 0001/01/01?*"

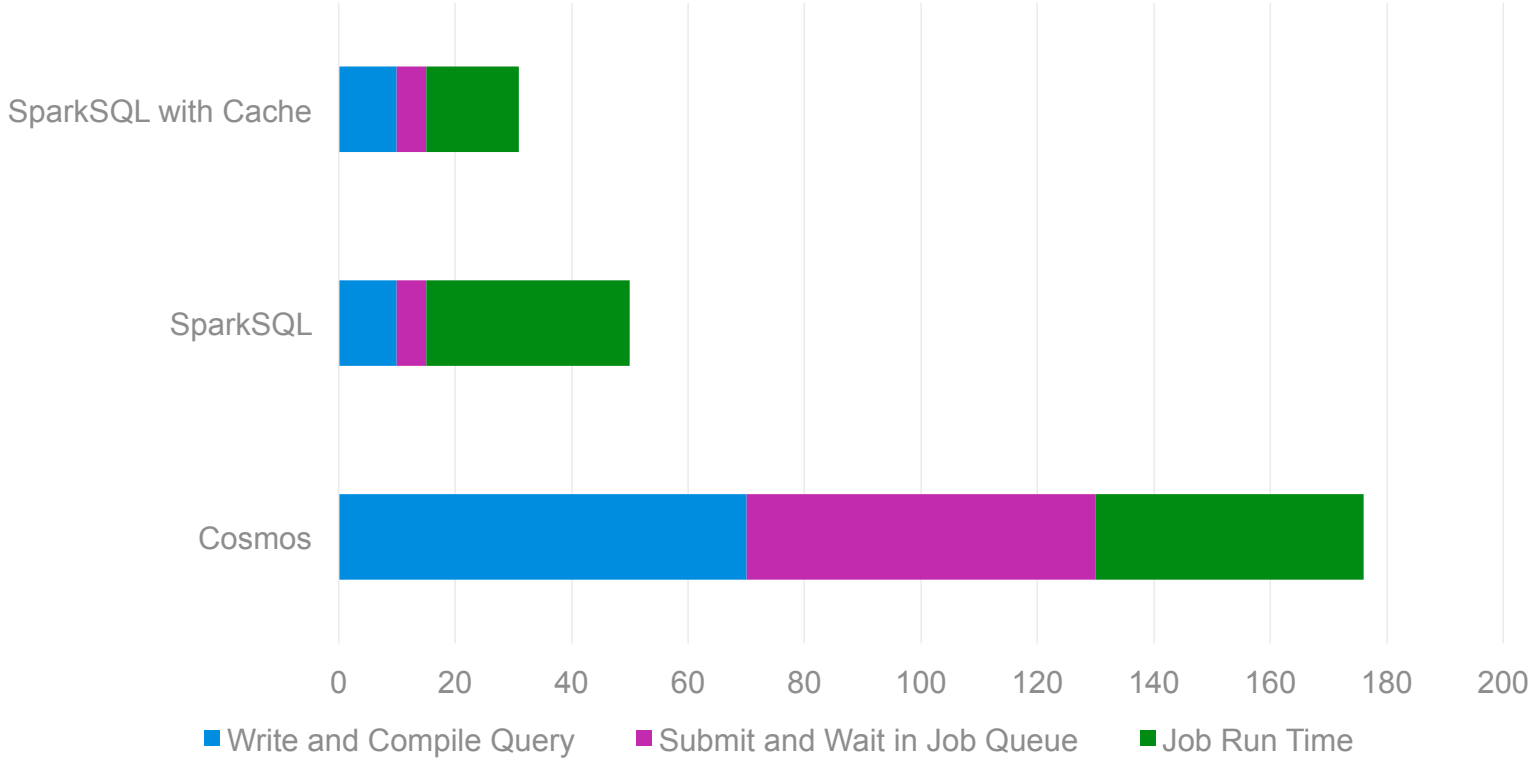
Is it by design or a bug?

71.2 TB of Office Data in Spark

What did Kuntao have to do?



Elapsed time



What did we learn in the process?

Learnings

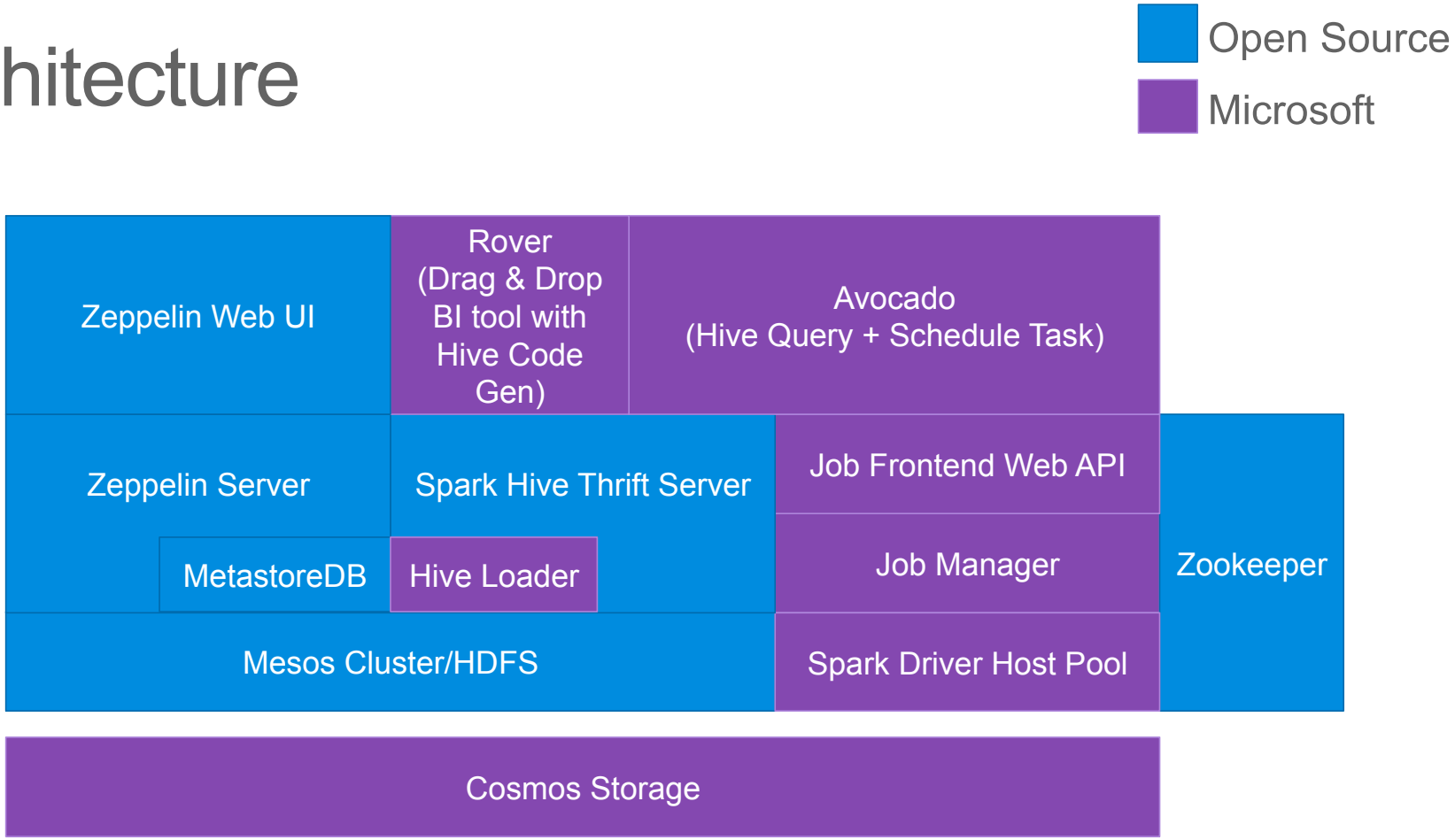
Enable self-service

Minimize time to insight

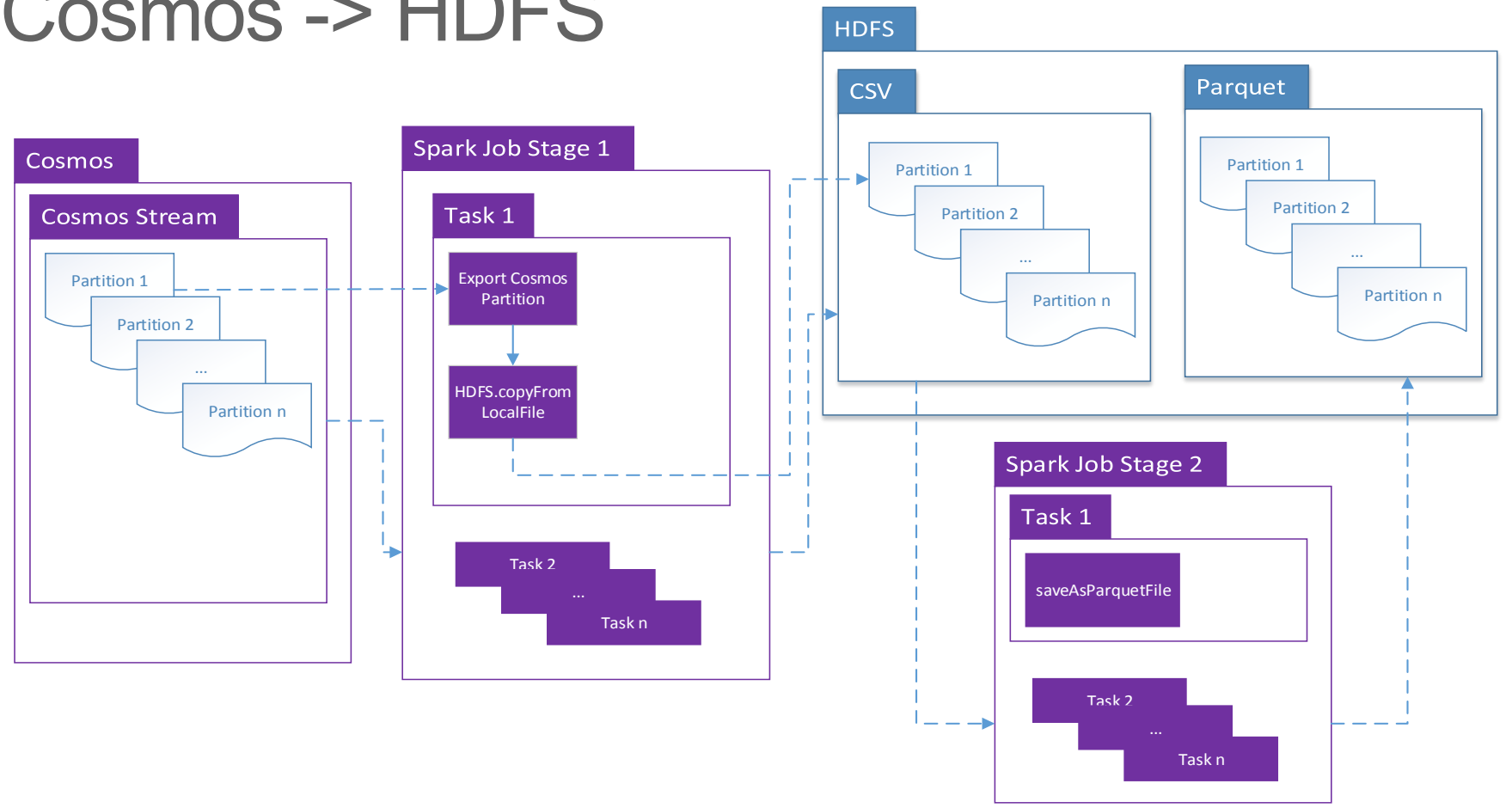
Notebooks at scale

How did we do it?

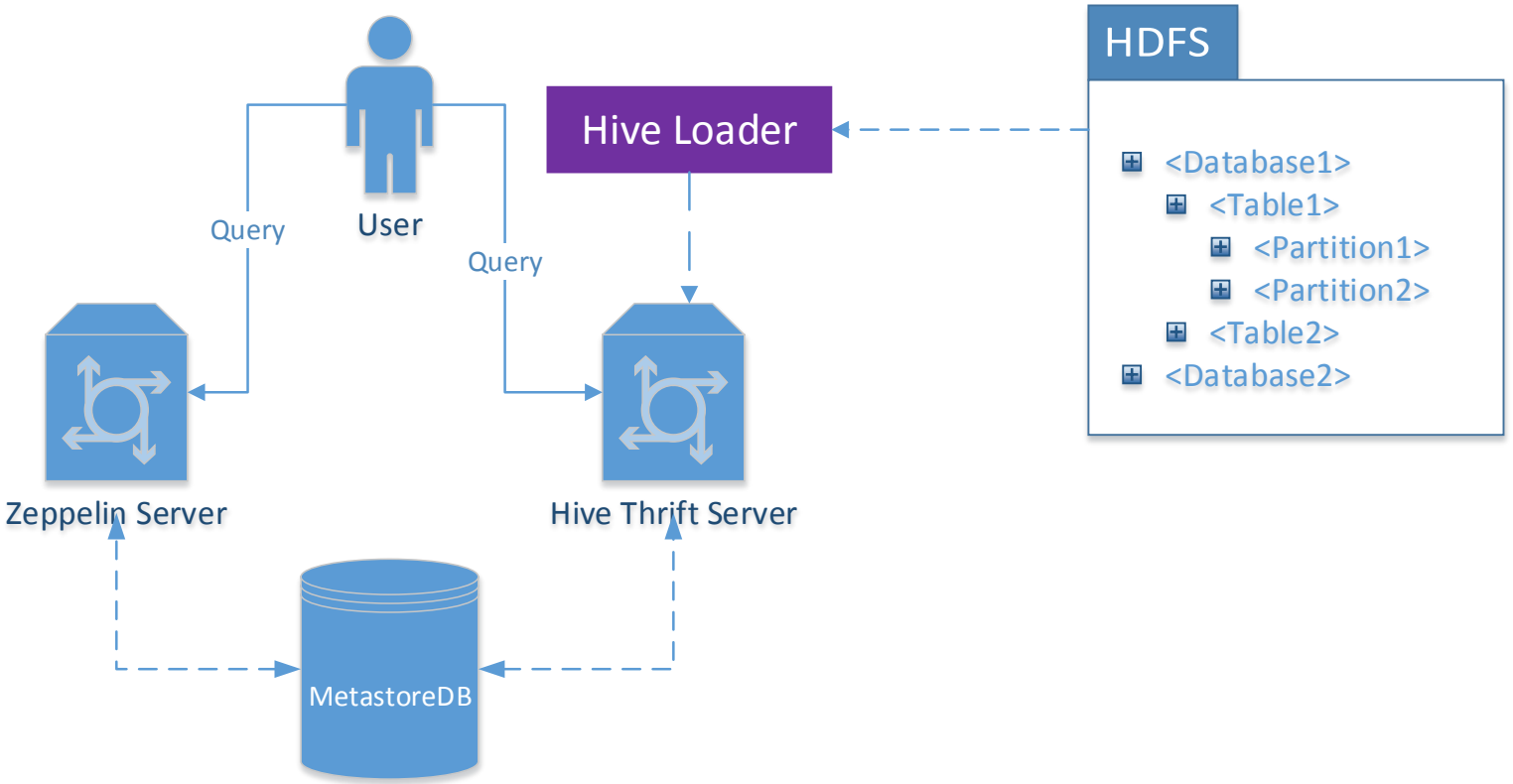
Architecture



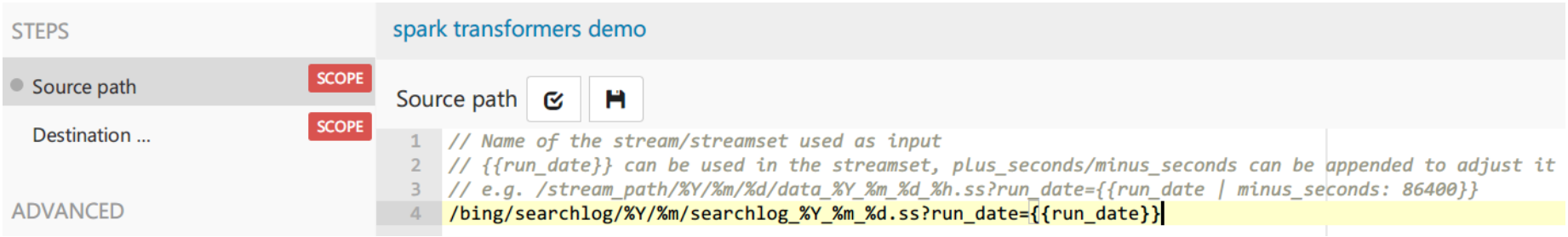
Cosmos -> HDFS

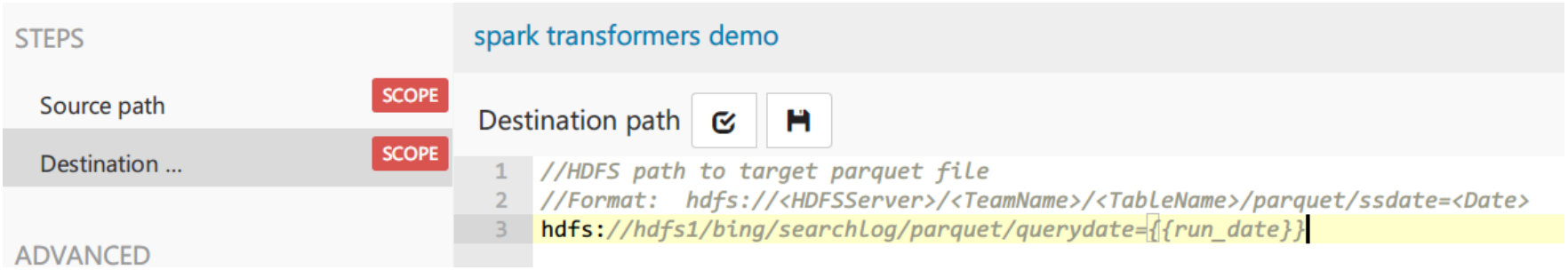

Cosmos
Cosmos Stream
Partition 1
Partition 2
...
Partition n
Spark Job Stage 1
Task 1
Export Cosmos Partition
HDFS.copyFrom LocalFile
Task 2
...
Task n
HDFS
CSV
Partition 1
Partition 2
...
Partition n
Parquet
Partition 1
Partition 2
...
Partition n
Spark Job Stage 2
Task 1
saveAsParquetFile
Task 2
...
Task n

Hive Loader


HDFS
Hive Loader
User
Query
Query
<Database1>
<Table1>
<Partition1>
<Partition2>
<Table2>
<Database2>
Zeppelin Server
Hive Thrift Server
MetastoreDB

User Experience in Avocado Task





User Experience in Avocado Task (Cont.)

Schedules + Add

	Status	Schedule ID	Interval	Parameters	Run Date	Recent Executions	
On Off	Stale	33855	1 day		2015-05-07 00:00:00		Backfill

Backfill Schedules

	Status	Schedule ID	Start Date	End Date	Parameters	Run Date	Recent Executions	
On Off	Running...	35010	2015-04-27 00:00:00	2015-04-27 00:00:00		2015-04-27 00:00:00		
On Off	Running...	34859	2015-03-01 00:00:00	2015-03-31 00:00:00		2015-03-21 00:00:00		
On Off	Running...	34860	2015-02-06 00:00:00	2015-02-28 00:00:00		2015-02-27 00:00:00		
On Off	OK	34895	2015-04-01 00:00:00	2015-04-01 00:00:00		2015-04-02 00:00:00		

User Experience in Avocado Task (Cont.)

More Info

Jobid	**20150507-202344-Avocado Cosmos to Spark Service-2a94c979-3234-4f77-be24-590acf0eb9b9**
Status	**RUNNING**
Driverurl	**http://sparkjob/proxy/?serverName=CO3828:4040**
Stdout	**http://sparkjob/logs/stdout-20150507-202344-Avocado Cosmos to Spark Service-2a94c979-3234-4f77-be24-590acf0eb9b9**
Stderr	**http://sparkjob/logs/stderr-20150507-202344-Avocado Cosmos to Spark Service-2a94c979-3234-4f77-be24-590acf0eb9b9**

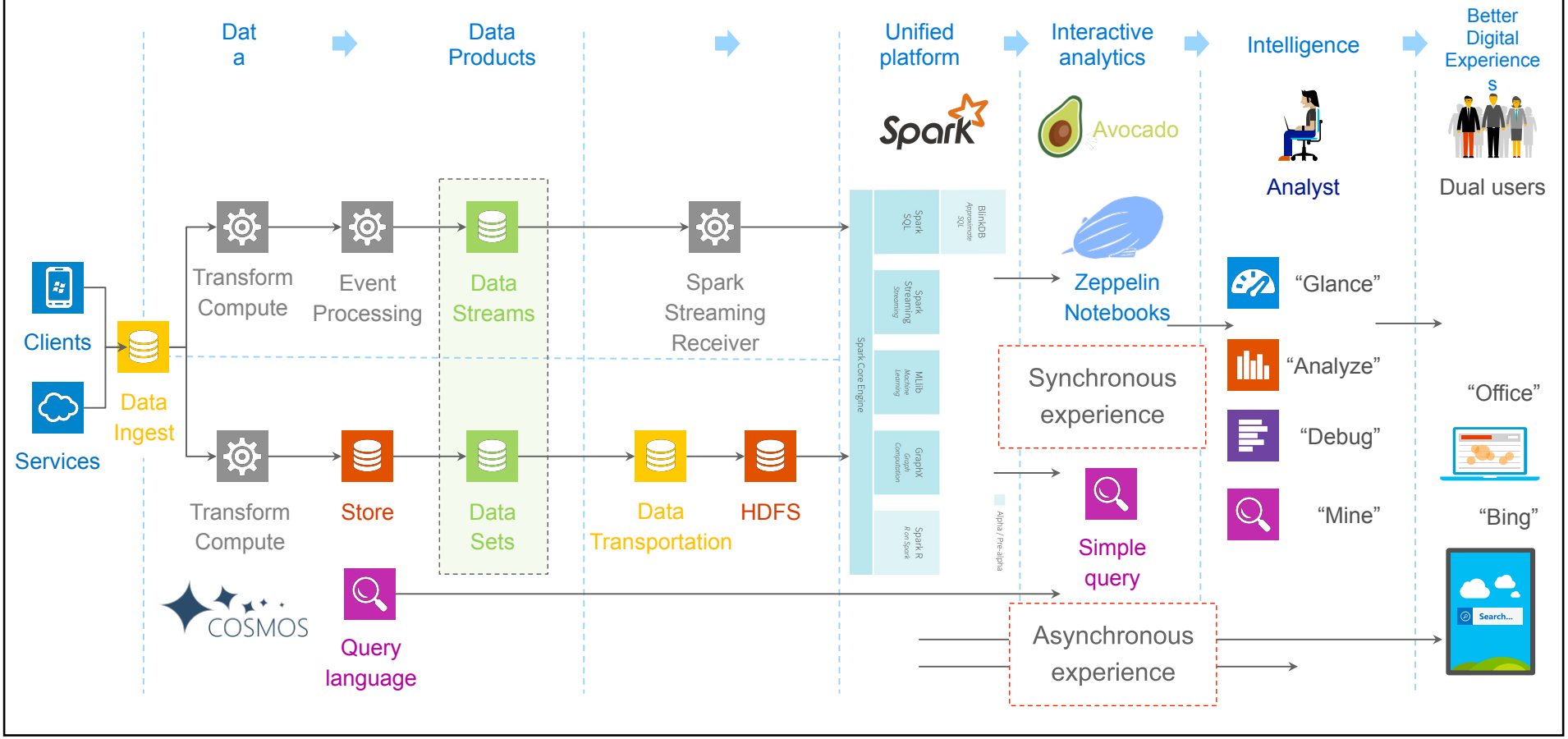
User Experience in Avocado Task (Cont.)

More Info

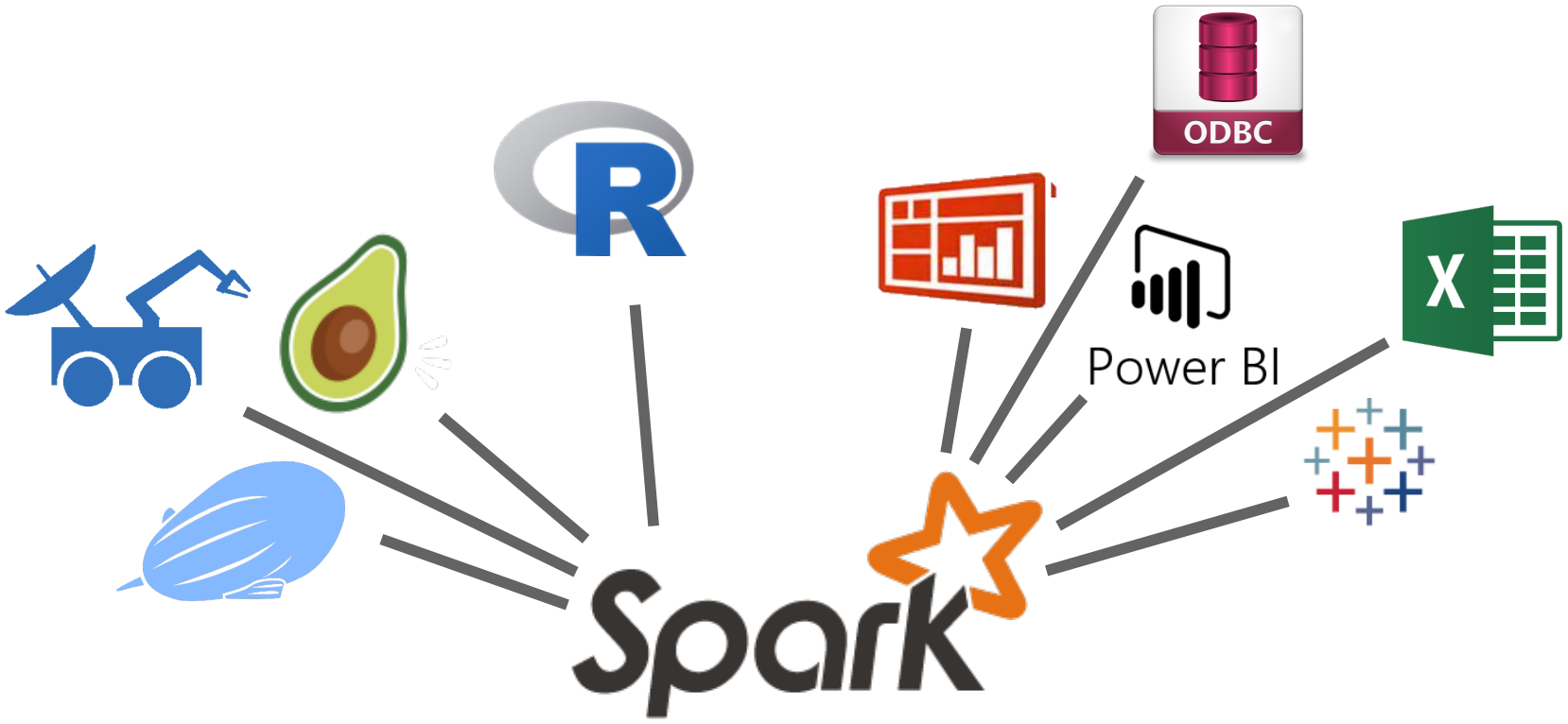
Jobid	20150507-184341-Avocado Cosmos to Spark Service-d170a297-e5e6-4bea-9972-888a71b167d0
Status	COMPLETED
Driverurl	Driver Url is unavailable after job finished, please check Jobhistoryurl
Stdout	http://sparkjob/logs/stdout-20150507-184341-Avocado Cosmos to Spark Service-d170a297-e5e6-4bea-9972-888a71b167d0
Stderr	http://sparkjob/logs/stderr-20150507-184341-Avocado Cosmos to Spark Service-d170a297-e5e6-4bea-9972-888a71b167d0
Jobhistoryurl	http://sparkjob/history/Framework-20150507_114940_1201_70221-52

What next?

Analytics as a Service in ASG



The #DataAnalystHub



Questions?

@JulienJTPierre
https://linkedin.com/in/julienjtpierre