Building a Data Warehouse for Business Analytics using Spark SQL

06/15/2015

Blagoy Kaloferov
Software Engineer

Today's talk

About me:

Blagoy Kaloferov

Big Data Software Engineer

About my company:

Edmunds.com is a car buying platform

18M+ unique visitors each month

Agenda

1. Introduction and Architecture
2. Building Next Gen DWH
3. Automating Ad Revenue using Spark SQL
4. Conclusion

Business Analytics at Edmunds.com

Divisions:

DWH Engineers

Business Analysts

Statistics team

Two major groups:

Map Reduce / Spark developers

Analysts with advanced SQL skills

Proposition

Spark SQL :

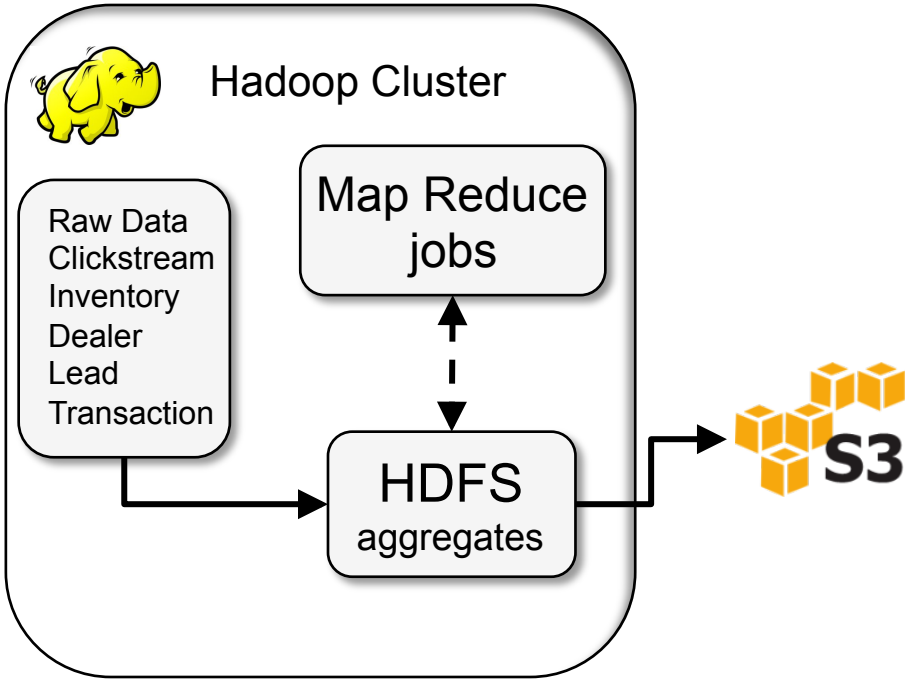
- Simplified ETL and enhanced Visualization tools
- Allows anyone in BA to quickly build new Data marts
- Enabled a scalable POC to Production process for our projects

Architecture for Analytics

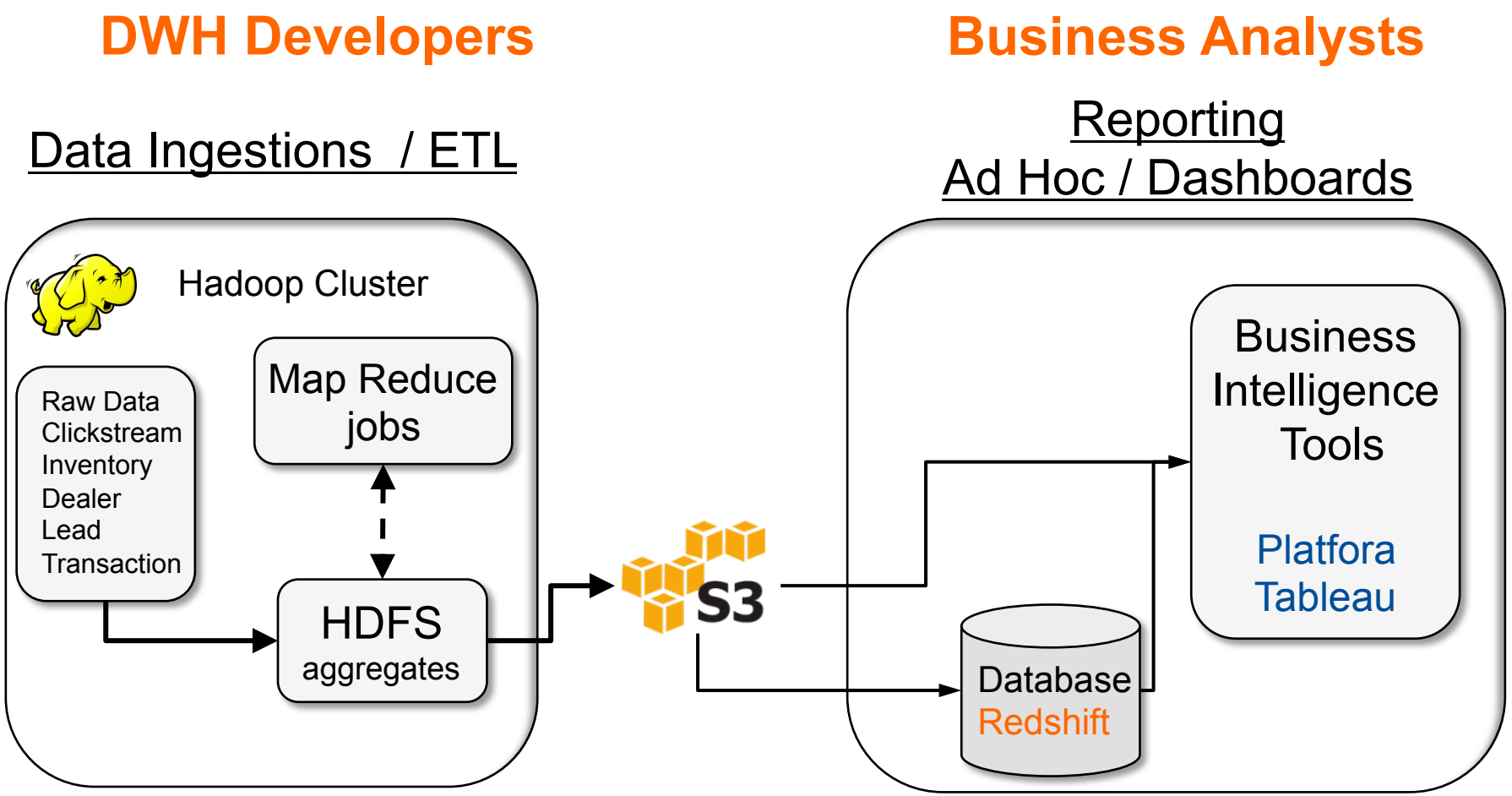
DWH Developers

Business Analyst

Data Ingestions / ETL



Architecture for Analytics



Architecture for Analytics

Databricks

Spark
Spark SQL

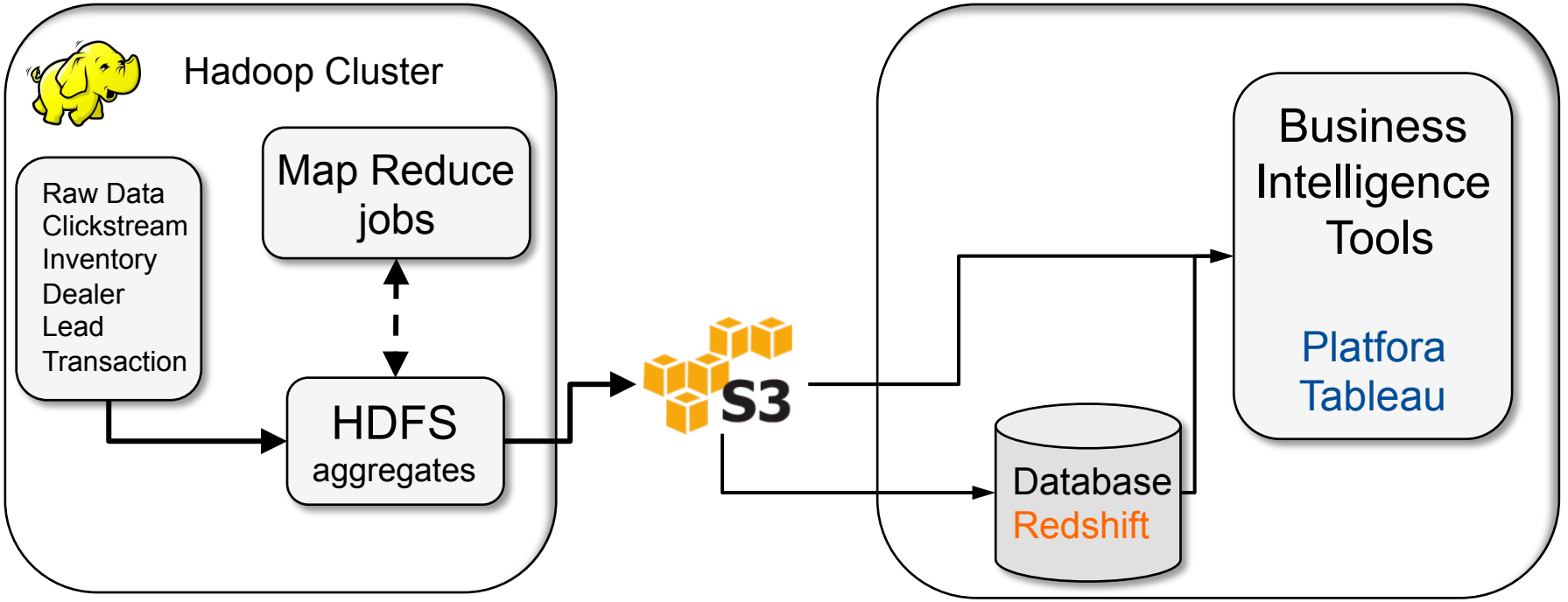
Spark

DWH Developers

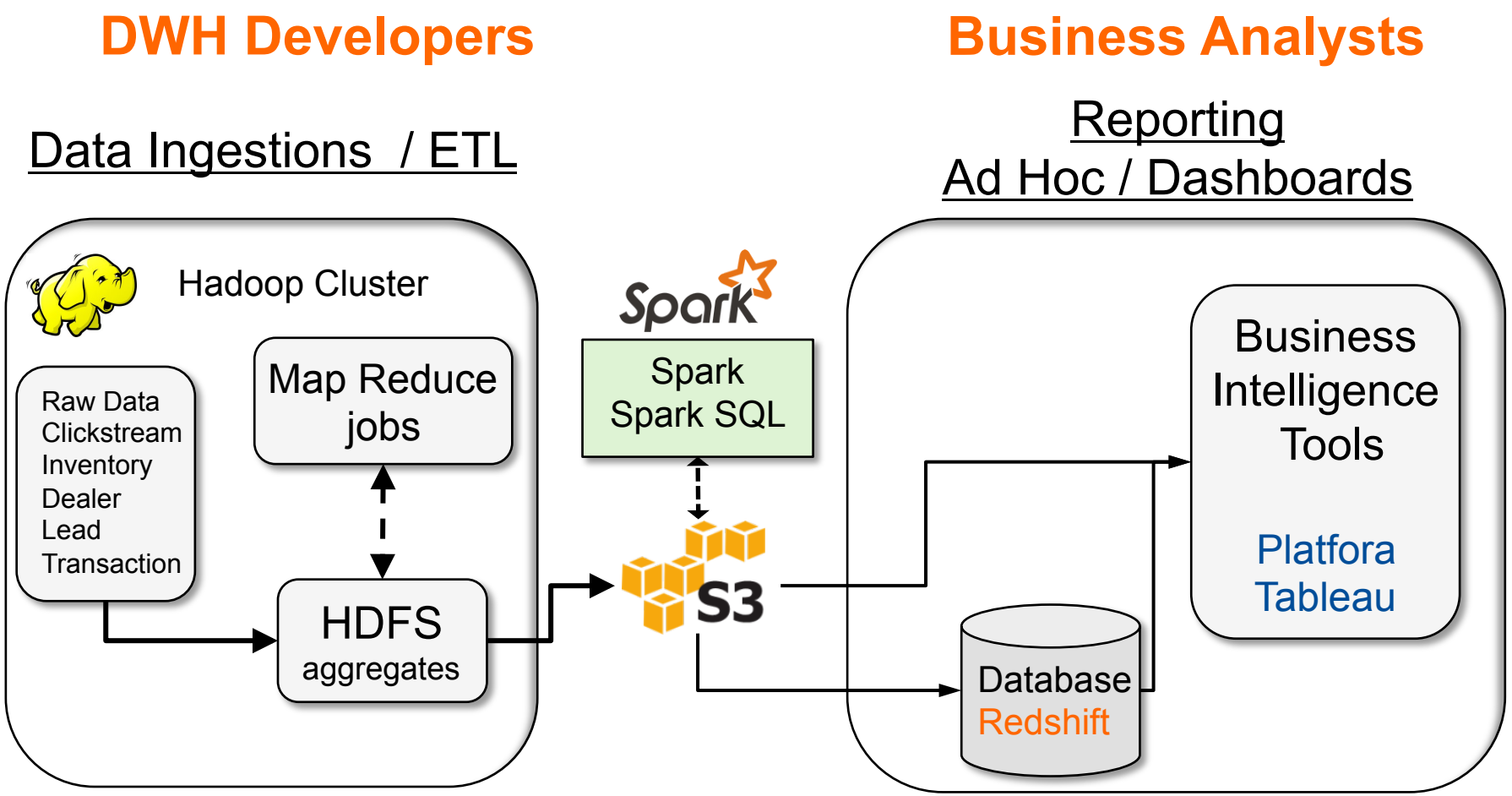
Business Analysts

Data Ingestions / ETL

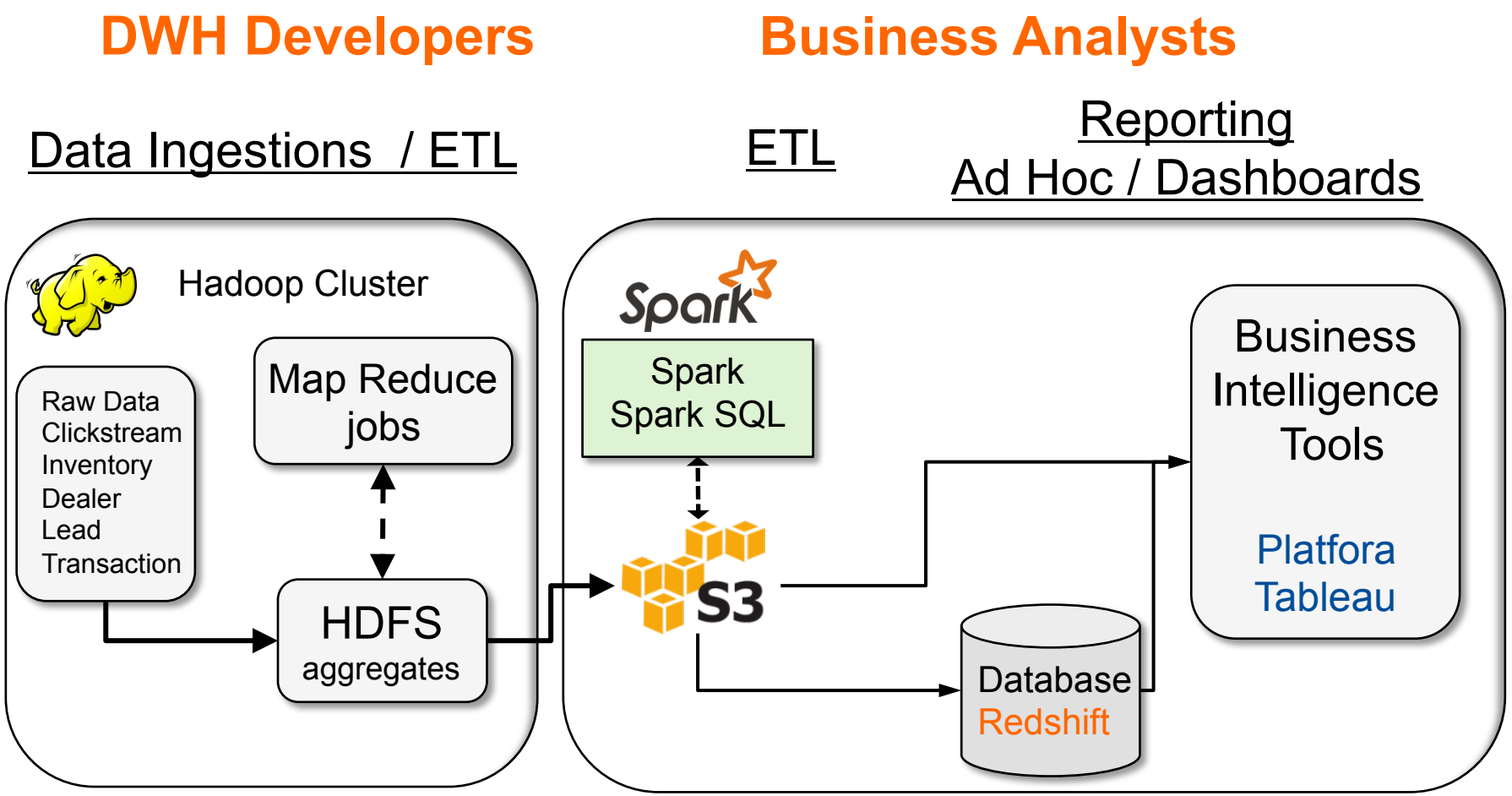
Reporting
Ad Hoc / Dashboards



Architecture for Analytics



Architecture for Analytics



Exposing S3 data via Spark SQL

Our approach:

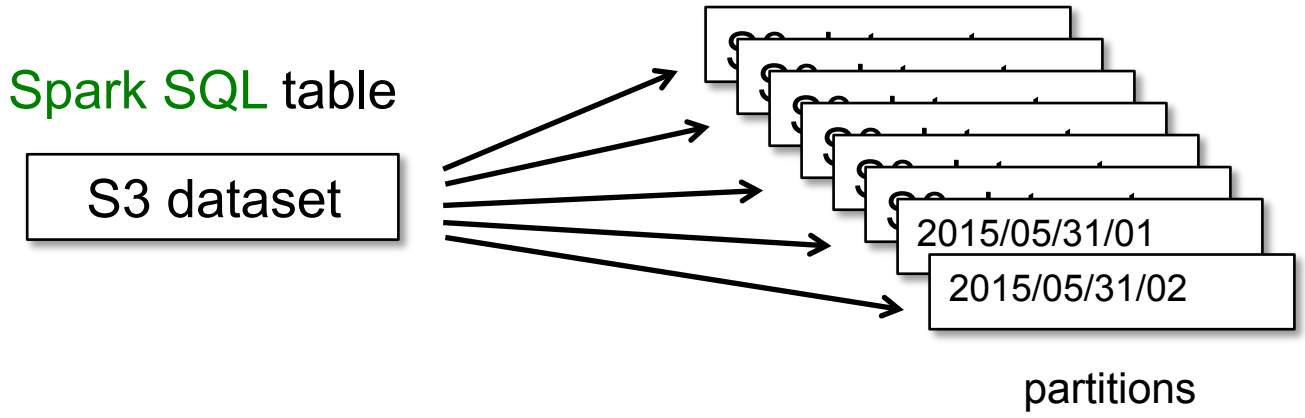
- Spark SQL tables similar to existing our Redshift tables
- Best fit for us are Hive tables pointing to S3 *delimited* data
- Exposed hundreds of Spark SQL tables

```
hiveContext.sql(
"""CREATE EXTERNAL TABLE clickstream
(
   clickstream_time TIMESTAMP,
   visitor_id STRING,
   visitor_key STRING,
   session_id STRING,
    ...
   device_make STRING
)
PARTITIONED BY (
  year STRING,
  month STRING,
  day STRING,
  hour STRING)
ROW FORMAT delimited
FIELDS TERMINATED BY '29'
STORED AS textfile
location 'root_location_S3/clickstream/';""" )
```

Spark SQL tables

Adding Latest Table Partitions

- S3 Datasets have thousands of directories:
  `( location /year/month/day/hour )`
- Every new S3 directory for each dataset has to be registered



Spark SQL tables

Utilities to for Spark SQL tables and S3:

- Register valid partitions of Spark SQL tables with spark Hive Metastore
- Create Last_X_ Days copy of any Spark SQL table in memory

Scheduled jobs:

- Registers latest available directories for all Spark SQL tables programmatically
- Updates Last_3_ Days of core datasets in memory

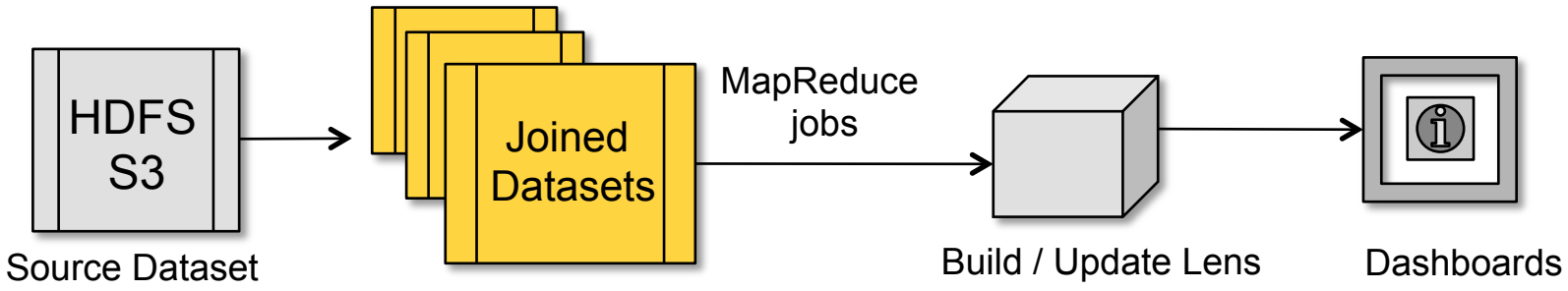
S3 and Spark SQL potential

- Now that all S3 data is easily accessible, there are a lot of opportunities !
- Anyone can ETL on prefixed aggregates and create new Data Marts

Platfora Dashboards Pipeline Optimization

platfora®

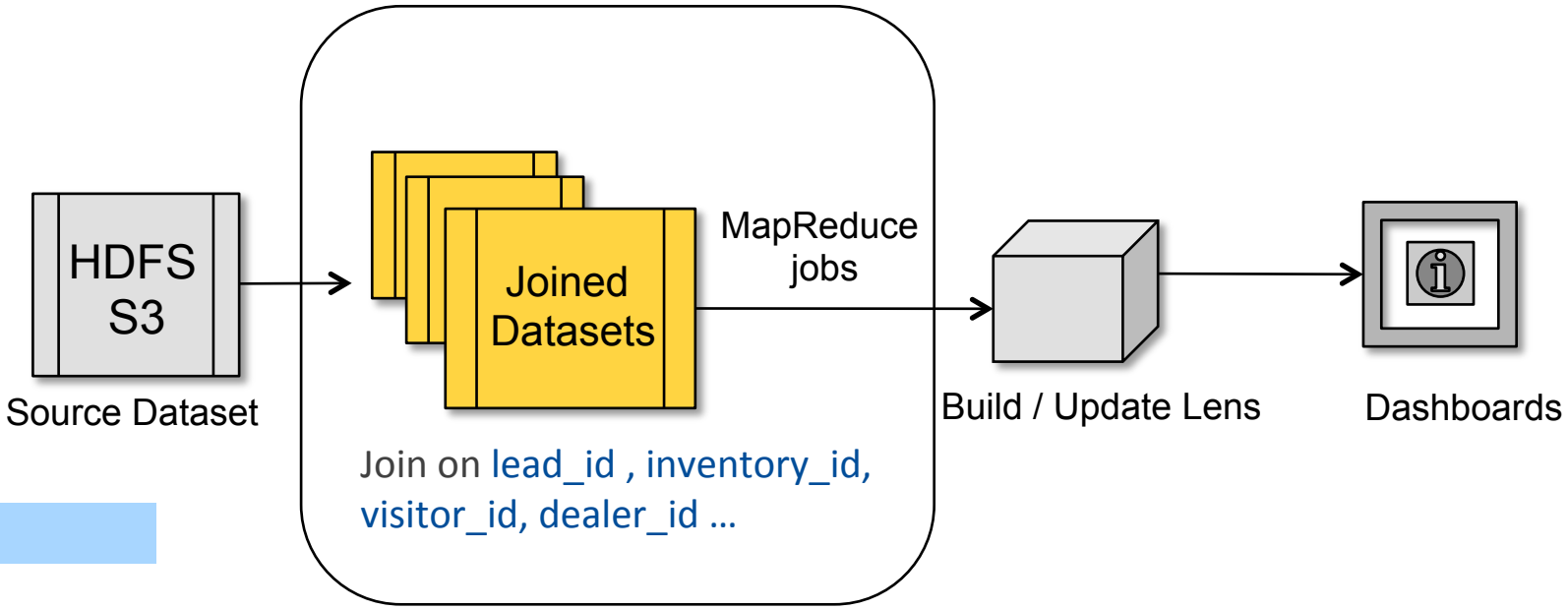
- Platfora is a Visualization Analytics Tool
- Provides More than *200 dashboards* for BA
  - Uses MapReduce to load aggregates



Platfora Dashboards Pipeline Optimization

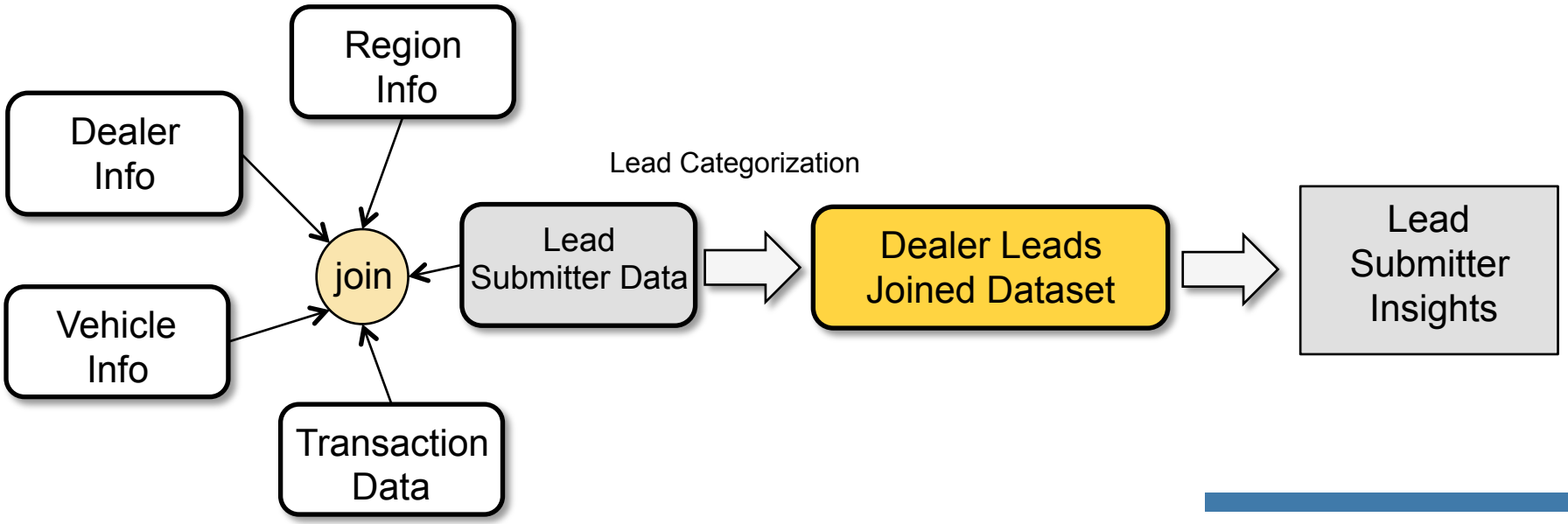
Limitations:

- We can not optimize the Platfora Map Reduce jobs
- Defined Data Marts not available elsewhere



Platfora Dealer Leads dataset Use Case

- Dealer Leads: Lead Submitter insights dataset
- More than 40 Visual Dashboards are using Dealer Leads



Optimizing Dealer Leads Dataset

Dealer Leads Platfora Dataset stats:

- 300+ attributes
- Usually takes 2-3 hours to build lens
- Scheduled to build daily

Optimizing Dealer Leads Dataset

How do we optimize it?

Optimizing Dealer Leads Dataset

How do we optimize it?

1. Have Spark SQL do the work!
   - All required datasets are exposed as Spark SQL tables
   - Add new useful attributes

2. Make the ETL easy for anyone in Business Analytics to do it themselves
   - Provide utilities and UDF's so that aggregated data can be exposed to Visualization tools

Dealer Leads Using Spark SQL Demo

Dealer Leads Data Mart using Spark SQL Demo

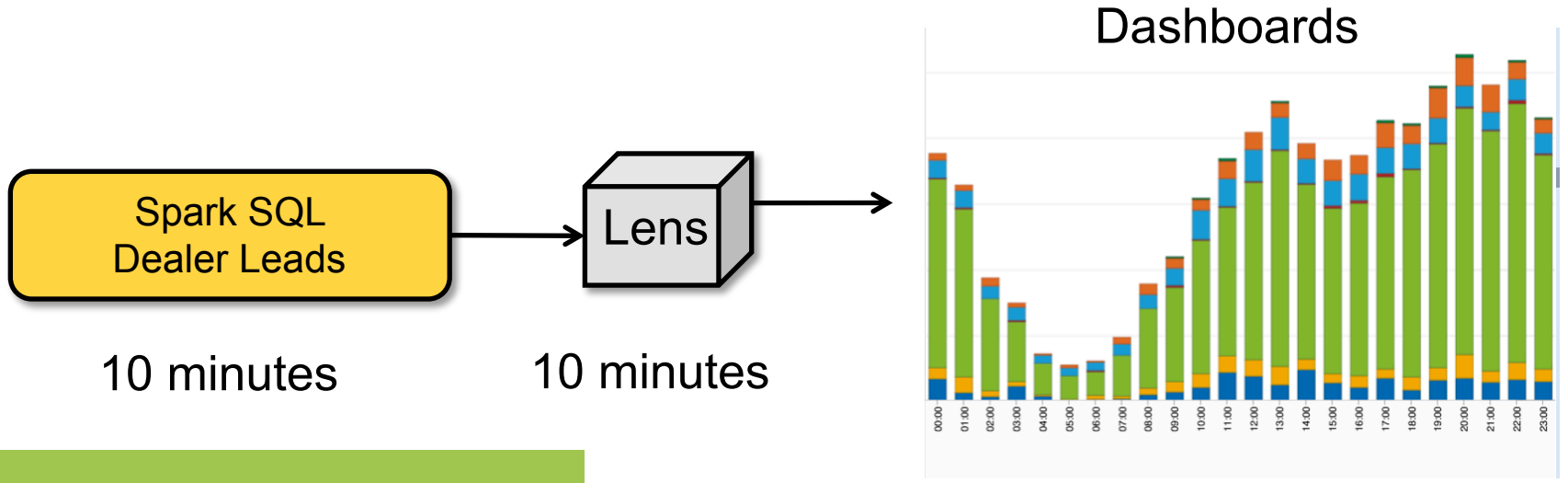
Expose all original 300+ attributes

Enhance: Join with site_traffic

aggregate_traffic_spark_sql

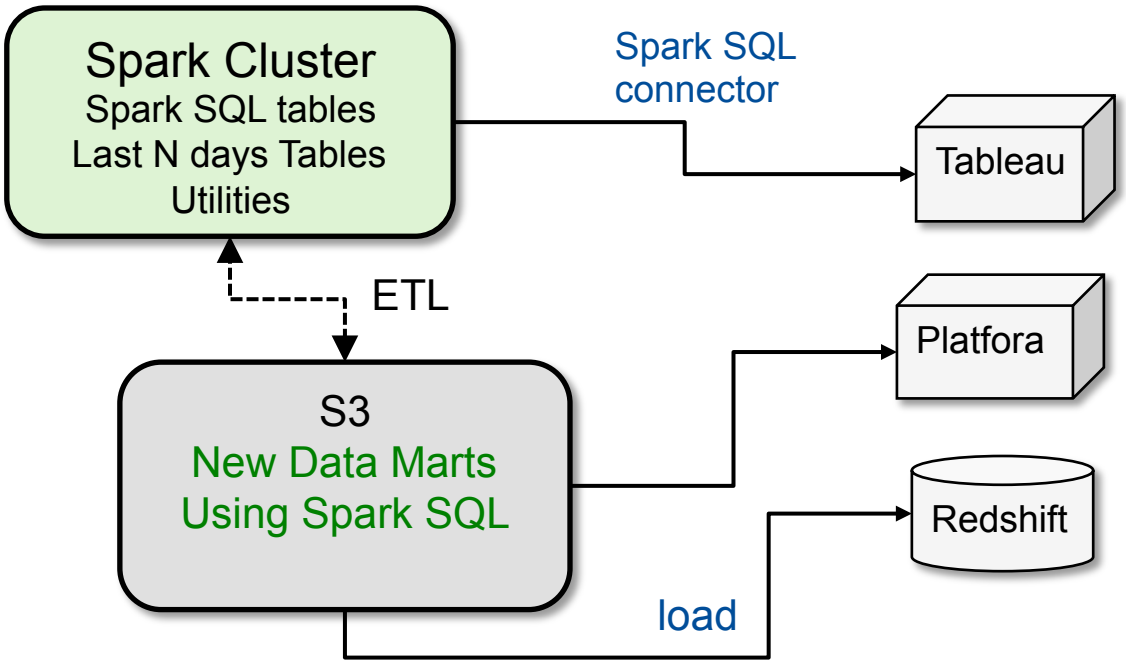
Dealer Leads Using Spark SQL results

- Spark SQL aggregation in 10 minutes.
  - Adds dimension attributes that were not available before
- Platfora does not need to join aggregates
- Significantly reduced latency
  - Dashboard refreshed every 2 hours instead of once per day.



ETL and Visualization takeaway

- Now anyone in BA can perform and support ETL on their own
- New Data marts can be exported to RDBMS



POC with Spark SQL vision

Usual POC process

- Business Analyst Project Prototype in SQL
  - Not scalable. Takes Ad Hoc resources from RDBMS
- SQL to Map Reduce
  - Transition from two very different frameworks
  - MR do not always fit complicated business logic.
  - Supported only by Developers

POC with Spark SQL vision

new POC process using Spark

- A developer and BA can work together on the same platform and collaborate using Spark
  - Its scalable
  - No need to switch frameworks when productionalizing

Ad Revenue Billing Use Case

Introduction

OEM Advertising on website

Definitions:

Impression, CPM

Line Item, Order

Ad Revenue Billing Use Case

Introduction

OEM Advertising on website

Ad Revenue computed at the end of the month using OEM provided impression data

Ad Revenue End of Month billing

Impressions served * *CPM* != actual revenue

- There are billing adjustment rules!
  - Each OEM has a set of unique rules that determine the actual revenue.
  - Adjusting revenue numbers requires manual user inputs from OEM's Account Manager

Line Item adjustments' examples

Box representing each example

- Line Item groupings

ORIGINAL Line Item | CPM | impressions | attributes

Line Item adjustments' examples

Box representing each example

- Line Item groupings

ORIGINAL Line Item | CPM | impressions | attributes

SUPPORT Line Item | CPM | impressions

Line Item adjustments' examples

Box representing each example

- Line Item groupings

Combine data

MERGED | CPM | NEW_impressions | attributes

Line Item adjustments' examples

Box representing each example

- Line Item groupings

MERGED | CPM | NEW_impressions | attributes

- Capping / Adjustments

Line Item | impressions | Contract

impressions_served > Contract ?

Line Item adjustments' examples

Box representing each example

- Line Item groupings

MERGED | CPM | NEW_impressions | attributes

- Capping / Adjustments

Line Item | CAPPED_impressions | Contract

impressions_served > Contract ?

Cap impression!

Line Item adjustments' examples

Box representing each example

- Line Item groupings

MERGED | CPM | NEW_impressions | attributes

- Capping / Adjustments

Line Item | impressions | Contract

impressions_served > (X% * Contract) ?

Line Item adjustments' examples

Box representing each example

- Line Item groupings

MERGED | CPM | NEW_impressions | attributes

- Capping / Adjustments

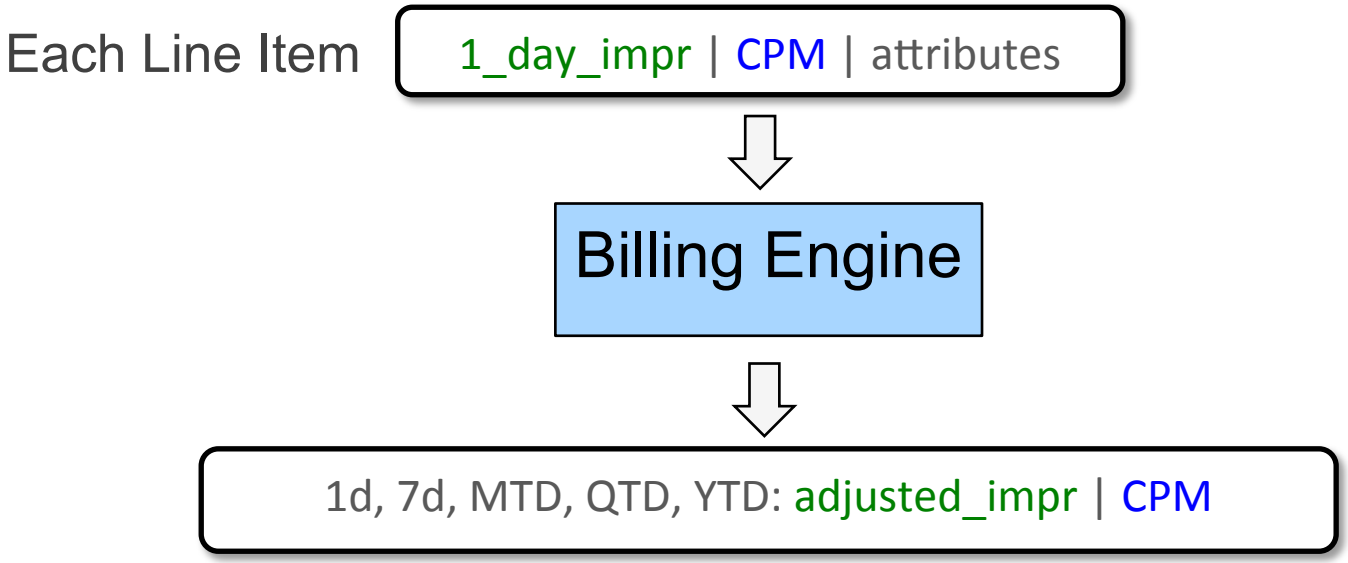
Line Item | ADJUSTED_impressions | Contract

impressions_served > (X% * Contract) ?

Adjust impression!

Can we automate ad revenue calculation?

Process Vision:



Impressions served * *CPM* = actual revenue

Can we automate ad revenue calculation?

Automation Challenges

- Many rules , user defined inputs, the logic changes
- Need for scalable unified platform
- Need for tight collaboration between OEM team, Business Analysts and DWH developers

Billing Rules Modeling Project

How do we develop it?

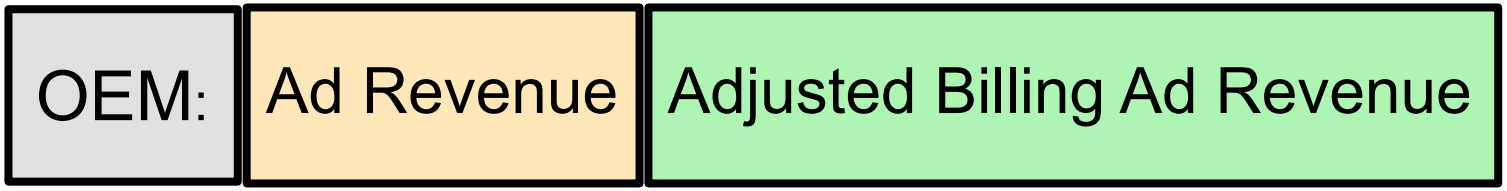
Billing Rules Modeling Project

How do we develop it?

Spark + Spark SQL approach

BA + Developers + OEM Account Team collaboration

Goal is an **Ad Performance Dashboard**



Project Architecture in Spark

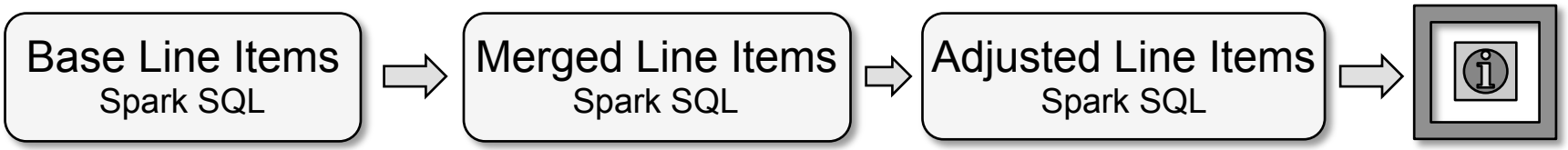
Each Line Item | 1_day_impr \| CPM \| attributes | = Spark SQL Row

Sum / Transform / Join Rows

Processing separated in phases where input / outputs are Spark SQL tables

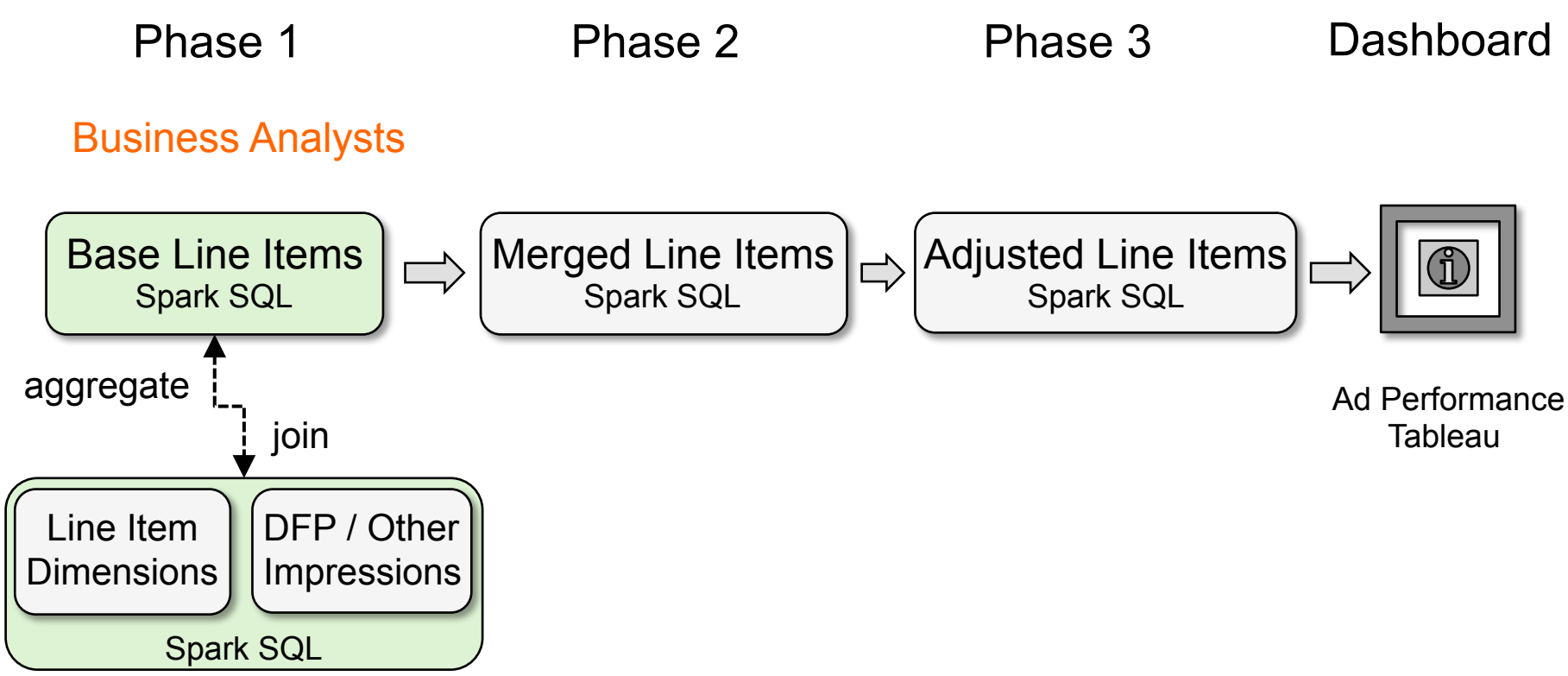
Project Architecture in Spark

Phase 1 | Phase 2 | Phase 3 | Dashboard

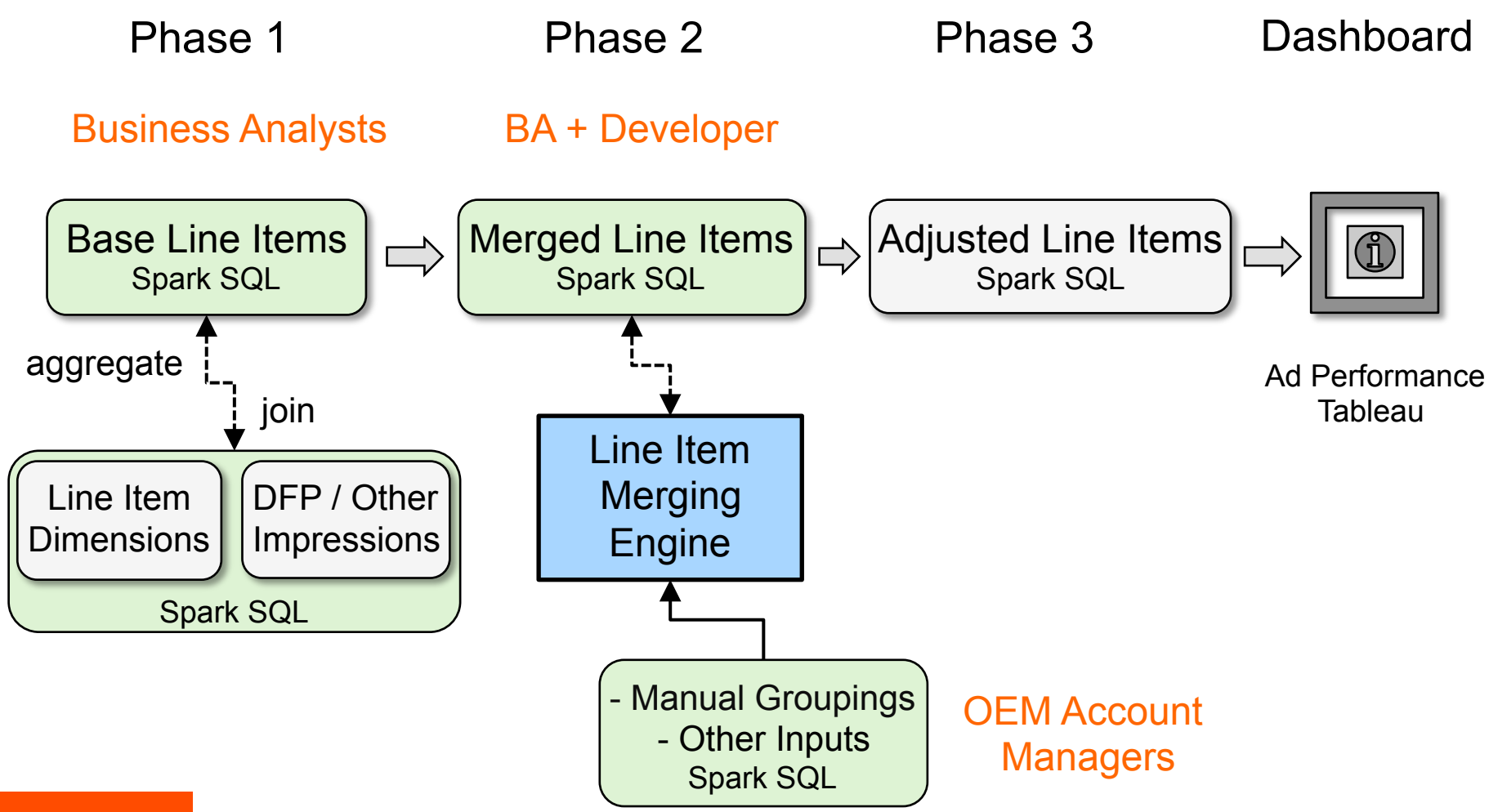


Ad Performance
Tableau

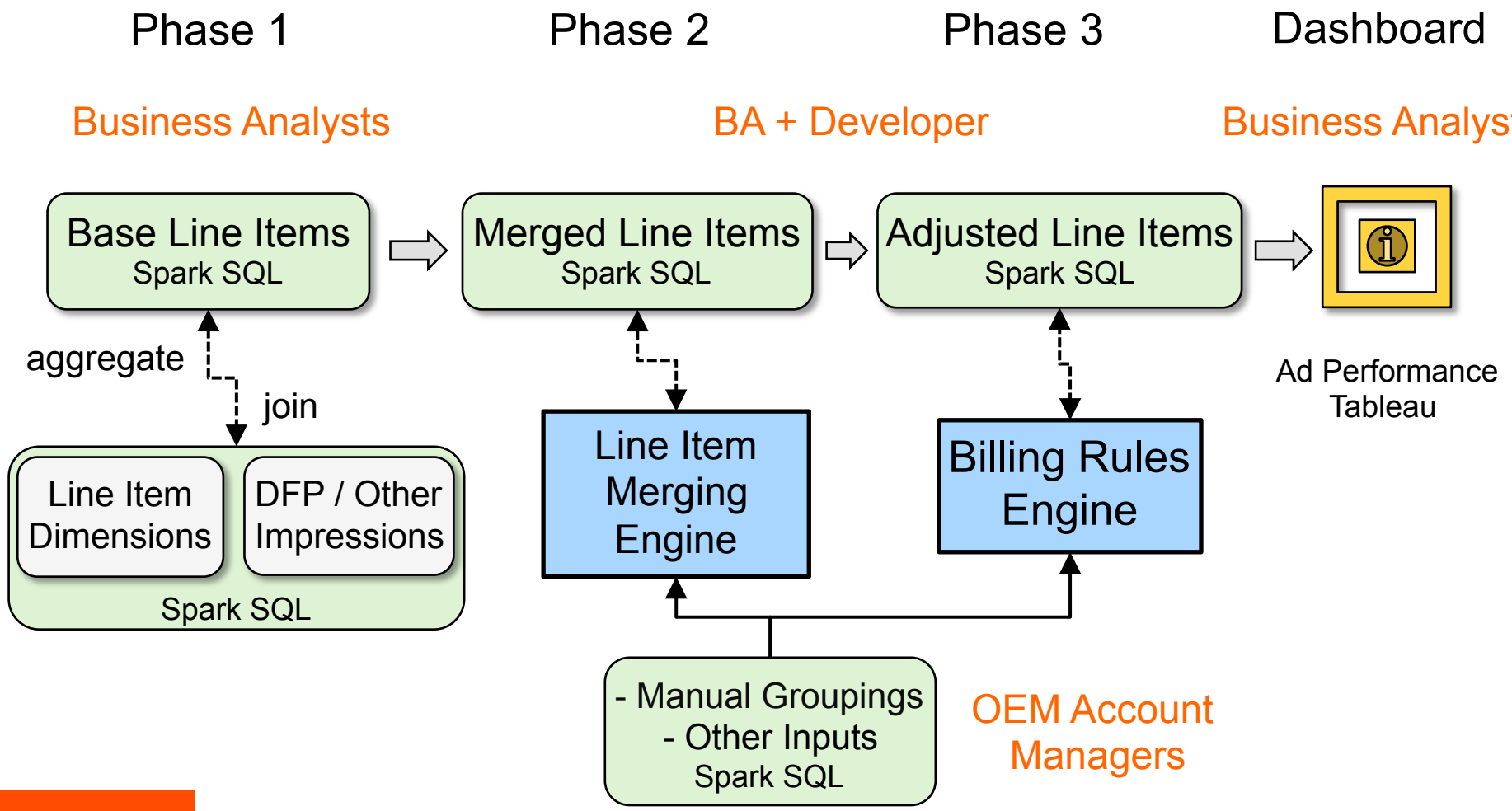
Project Architecture in Spark



Project Architecture in Spark



Project Architecture in Spark



Billing Rules Modeling Achievements

- Increased accuracy of revenue forecasts for BA
- Cost savings by not having a dedicated team doing manual adjustments
- Monitor ad delivery rate for orders
  - Allows us to detect abnormalities in ad serving
- Collaboration between BA and DWH Developers

Questions?

Thank you!

Blagoy Kaloferov

bkaloferov@edmunds.com