OpenWorld 2018 SparklineData: Data Lake Analytics At Scale

Apache Spark Native OLAP Queries

Sridhara Sabbella
Director Products
Big Data, Oracle Cloud
October 23rd, 2018

ORACLE OPEN WORLD

ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Program Agenda

1. Introduction
2. SNAP Overview
3. Why SNAP?
4. Examples
5. Summary

Introduction

The need for Fast Queries

ORACLE

Fast Analytics on Data Lakes – Key Drivers

Enterprise Data consolidated in ***Cloud Data lakes***

- HDFS, Object storage

SQL on Hadoop is not enough for ***B.I and business users***

Need ***fast*** visualization directly on datalakes

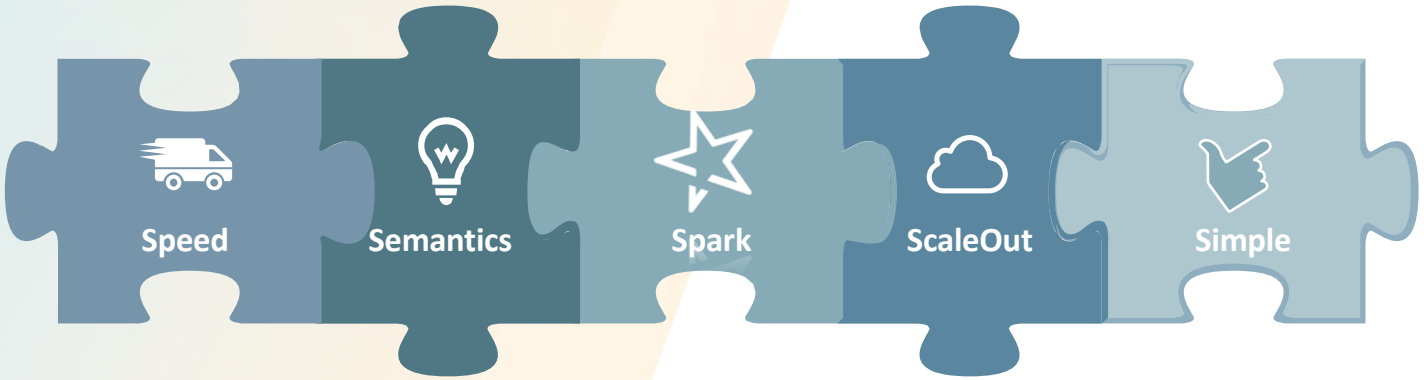
Need ***open*** data management platforms (Spark)

SNAP Overview

Fast OLAP style queries on Spark

ORACLE

SNAP : 5 key features



Fast In-Memory OLAP Index

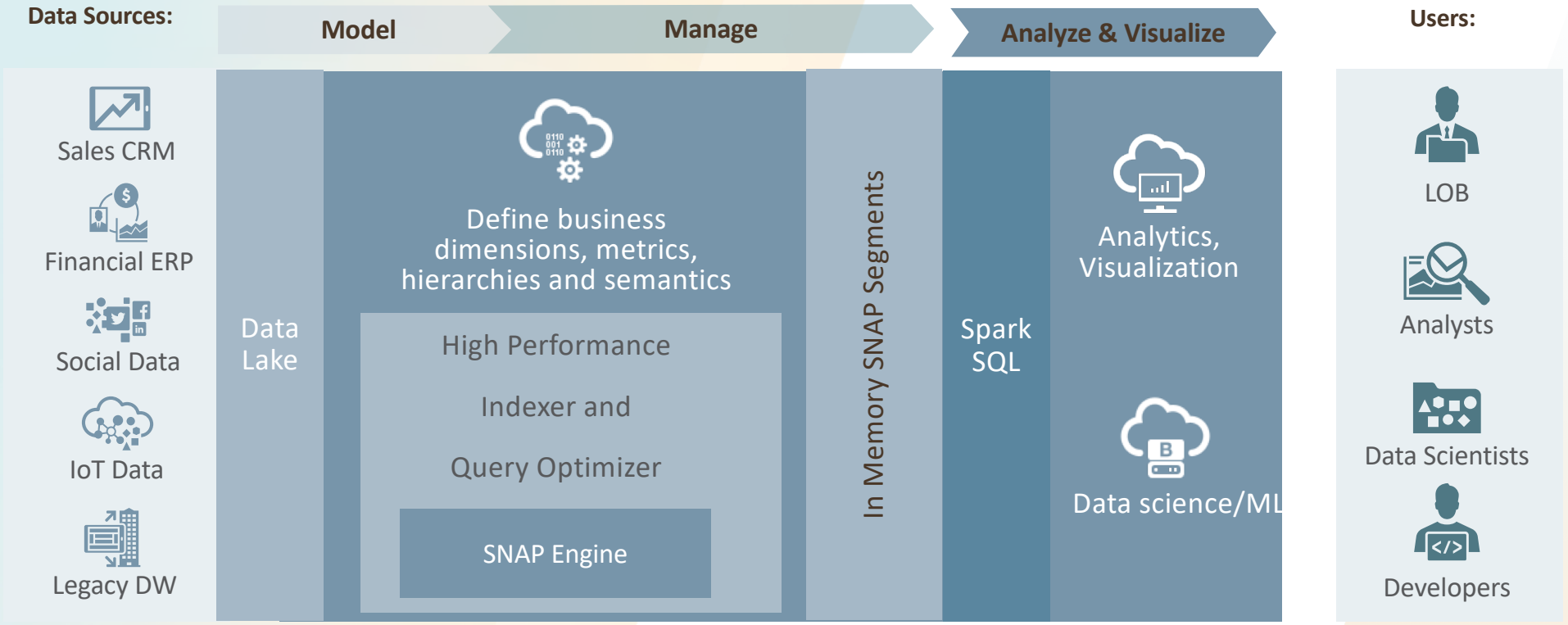
Query planning and rewrites based on business semantics

Fully built on Apache Spark /Open

Architected for the cloud - elastic and scale out

Just Object Storage and a Spark Cluster.

What is SNAP?


Data Sources:
Model
Manage
Analyze & Visualize
Users:
Sales CRM
Financial ERP
Social Data
IoT Data
Legacy DW
Data Lake
Define business dimensions, metrics, hierarchies and semantics
High Performance
Indexer and
Query Optimizer
SNAP Engine
In Memory SNAP Segments
Spark SQL
Analytics, Visualization
Data science/ML
LOB
Analysts
Data Scientists
Developers

Why SNAP?

Differentiation

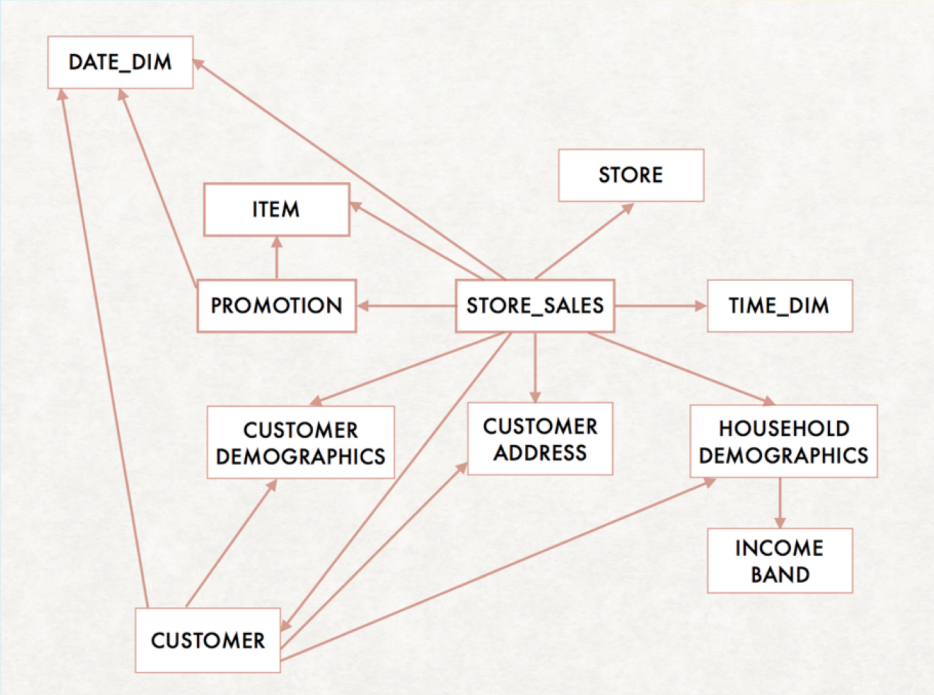
Why SNAP?

A typical business question is not a simple SQL Query

Business context examples –

- **Analysis on certain locations: ‘New York’, ‘Seattle’,... -the first 4 months of 2017**
- **Slice by customer demographics and income levels and channel**
- **Metrics – Simple and advanced**
- **Revenue, Profits and Units sold**
- **Ratios (share of profits based on different levels in a hierarchy)**
- **Apportioning /Allocation – Assign revenue to territories based on units sold and then calculate “Derived Revenue”**

A typical analysis on datasets



Facts capture events – Store sales

Sales of items, quantity sold etc

Several dimensions describe the fact

- Who bought it (customers/users, demographics, etc)
- What was bought (item)
- When was it bought (time)

Joins are not performant

Expressing business logic in SQL requires specialized skill sets – not business friendly

Tableau and other tools cannot handle several million row facts and joins without extracts

Answering business queries with SparklineSNAP

Queries on a business “QUBE”

Model commonly analyzed dimensions and metrics

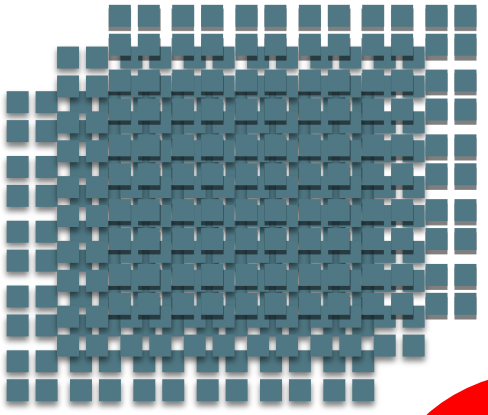
- channel name
- location hierarchy (country, state, city)
- Product hierarchy
- Revenue
- Quantity
- Ratios
- Time hierarchies- Calendar, Fiscal

A DSL in the future to express analytics queries

Much simpler query plans making use of business semantics to optimize execution plan

Makes use of Query optimization, In-memory indexing and more

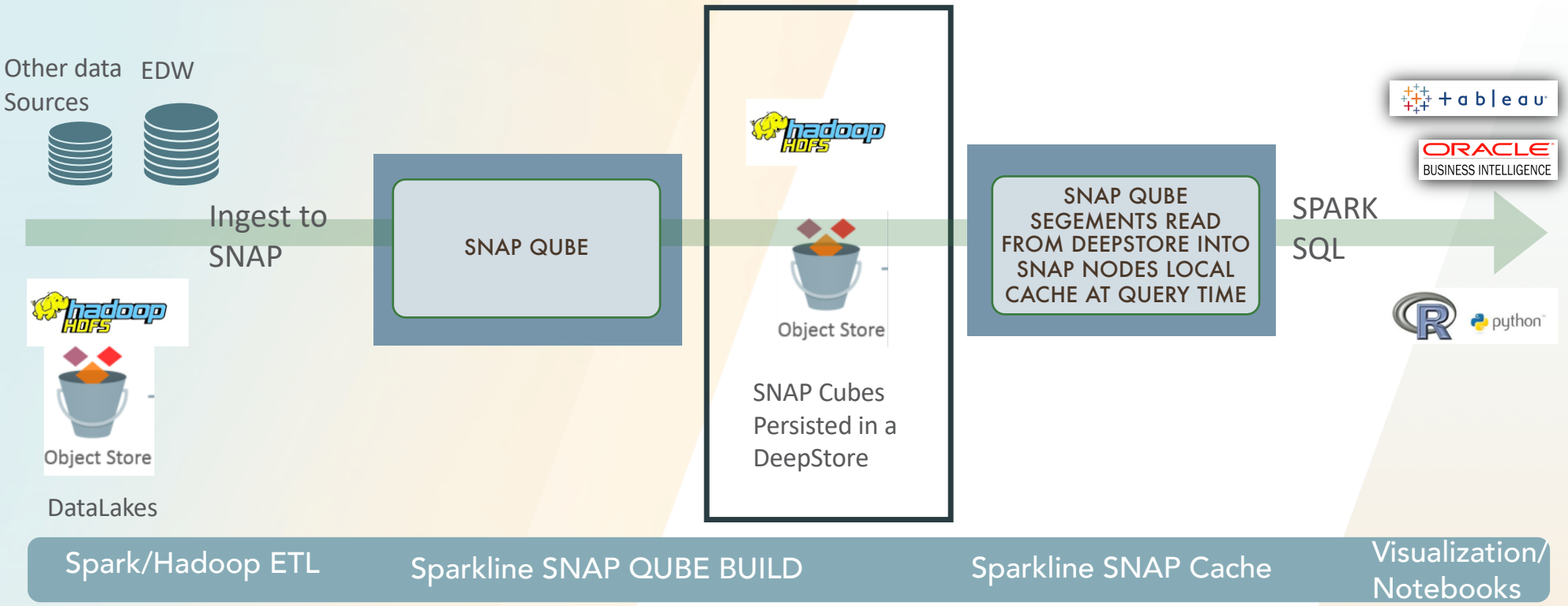
SNAP Multi-dimensional OLAP Index



No Pre-Aggregations

FAST and simple

SNAP QUBE – Flow



Architecture

SNAP Segments on local NVME cache(Distributed Spark worker nodes)

VM DENSE IO

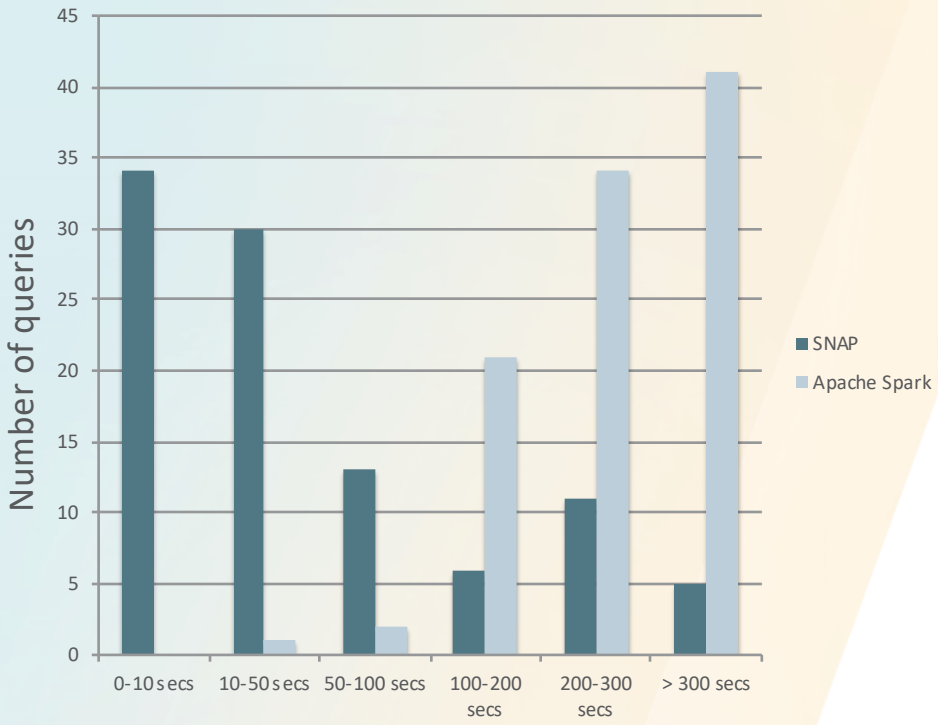
Oracle Data Lake and Compute

Qube Storage(on OCI Object Storage)

External Data Lakes

Source Data can reside anywhere

SNAP vs Apache Spark



SNAP Query times for the 99 queries are bunched on the left with most between 0-10 s

Spark Query times are bunched to the right with over 200s

For OLAP queries SNAP is up to 100x faster

Internal TPCDS test run on AWS/OCI

Examples

SNAP in production

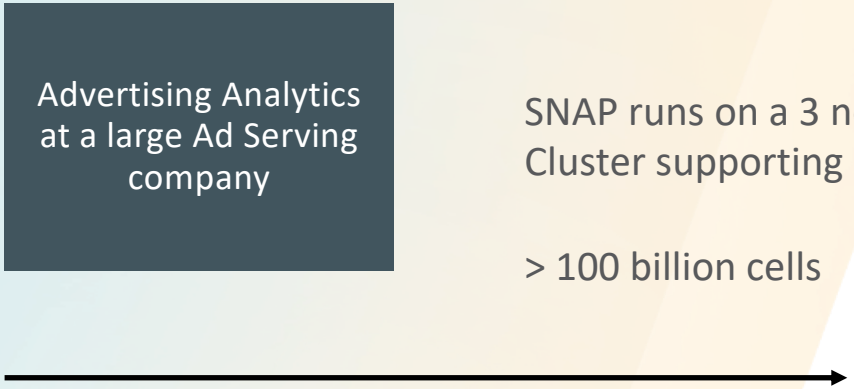
AD Tech use case

Query as a service

Advertising Analytics at a large Ad Serving company

SNAP runs on a 3 node standalone
Cluster supporting multiple billions of rows

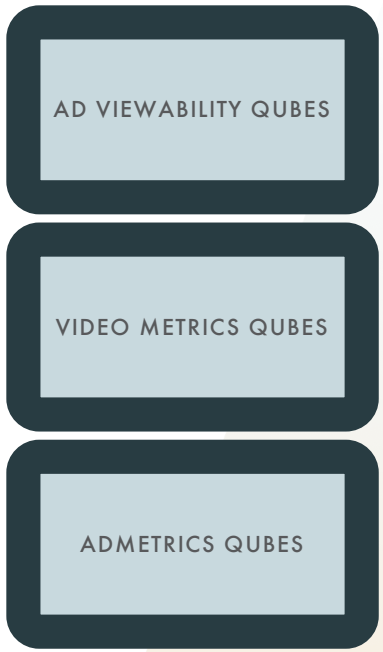
> 100 billion cells



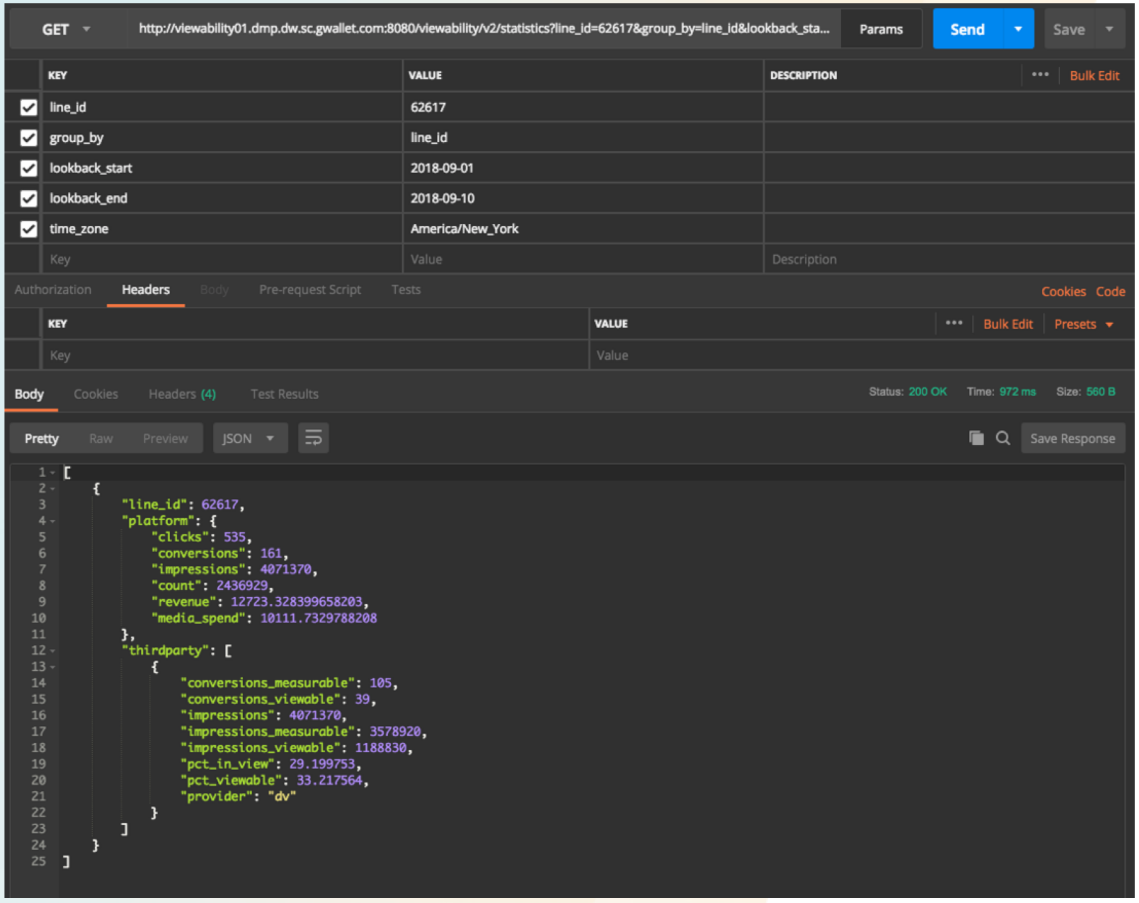
Multi- TB datamart

< 1 s response times

100s of concurrent sessions from web apps



Targeted Queries from APIs to Sparkline SNAP

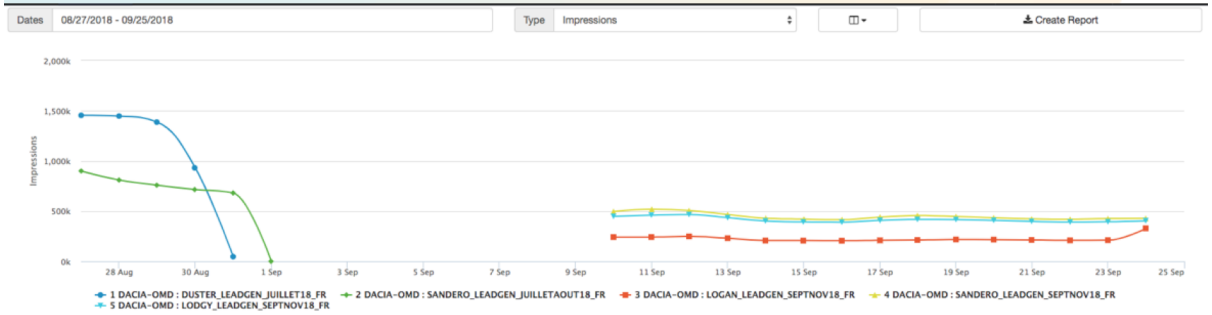


Programmatic API Queries to SNAP

Running 24x7 against SNAP QUBES

Roundtrip response time in milliseconds

B.I on SNAP for Ad Reporting



Campaign	Impressions	Clicks	CTR(%)	VCR(%)	Cost
DACIA-OMD : SANDERO_LEADGEN_SEPTNOV18_FR	6 732 851	2 009	0,030%	0,000%	3 819,00 €
DACIA-OMD : LODGY_LEADGEN_SEPTNOV18_FR	6 236 760	1 339	0,021%	0,000%	2 364,00 €
DACIA-OMD : DUSTER_LEADGEN_JUILLET18_FR	5 269 527	988	0,019%	0,000%	1 480,00 €
DACIA-OMD : SANDERO_LEADGEN_JUILLETAOUT18_FR	3 865 603	951	0,025%	0,000%	1 282,00 €
DACIA-OMD : LOGAN_LEADGEN_SEPTNOV18_FR	3 411 240	880	0,026%	0,000%	2 236,00 €
RENAULT-OMD : CAPTUR_LEADGEN_SEPTNOV18_FR	1 534 180	789	0,051%	0,000%	2 522,00 €
DACIA-OMD : DOKKER_LEADGEN_JUILLETAOUT18_FR	1 525 684	350	0,023%	0,000%	495,00 €
RENAULT-OMD : SCENIC_LEADGEN_SEPTNOV18_FR	1 124 054	489	0,044%	0,000%	1 736,00 €
DACIA-OMD : LODGY_LEADGEN_JUILLETAOUT18_FR	947 756	136	0,014%	0,000%	261,00 €
DACIA-OMD : LOGAN_LEADGEN_JUILLETAOUT18_FR	211 017	45	0,021%	0,000%	76,00 €
	30 858 672	**7 976**	**0,026%**	**0,000%**	**16 276,00 €**

Example query and some stats:

```
SELECT SUM(impressions) as impressions, SUM(clicks) as clicks, S
UM(conversions) as conversions,

 SUM(view_through_conversions) as view_through_conversions,
 SUM(video_view_start) as video_view_start,

 SUM(video_view_1) as video_view_1, SUM(video_view_2) as vi
deo_view_2,

 SUM(video_view_3) as video_view_3, SUM(video_view_end) as
video_view_end,

 SUM(revenue) as revenue, campaign_id

FROM moat_daily_snap_fixed

WHERE daily_snap_fixed.advertiser_id IN (5738)

 AND (dt BETWEEN '20180827' AND '20180925')

GROUP BY campaign_id
```

Roundtrip 1.22s
1 Billion rows a month
QUBE has 13 months
of data

ERP Datalake use case

FINANCE DATA LAKE

ERP data analytics at a large Fortune 100 conglomerate

SNAP runs on a shared HDP YARN managed cluster supporting

500 Billion cells
(500 million rows X 1000 columns)

Migrated from → Sparklinedata

Greenplum

Multiple TB datasets
Across 25+ tables in star schema

< 2s response times for joins/hierarchies from OBIEE

- GENERAL LEDGER QUBES
- TRAVEL/EXPENSE REPORT QUBES
- PAYABLES QUBES
- ITEMS QUBES

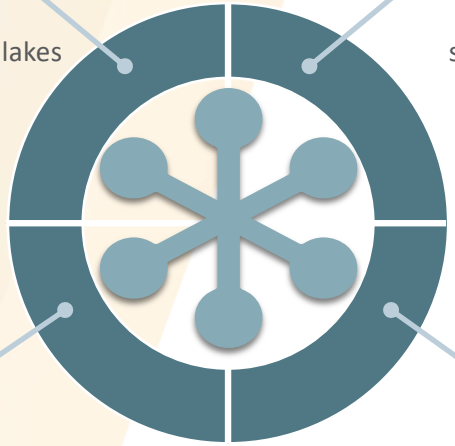
Summary

Business focused

Model your business domain data and define semantics as cubes for Enterprise B.I on data lakes

FAST @ SCALE

Up to **100x faster** than comparable query solutions on data lakes. Elastic Scale out on the Oracle cloud



Enterprise ready

Fast and Simple deployment – It is just a Spark Cluster! Plugs into to your existing B.I products such as OAC/Tableau

High ROI

Eliminate expensive legacy datamarts and extracts and standardize on Spark based Interactive Analytics