Big Data Analytics Platform @ Nokia

Selecting the Right Tool for the Right Workload

Yekesa Kosuru
Nokia

Location & Commerce

Strata + Hadoop World NY - Oct 25, 2012

Agenda

- **Big Data Analytics Platform @Nokia**
  - **Who we are**
  - **Use case data flows**
  - **Big data platform**
  - **Big data challenges**
- **Selecting the Right Tool for the Right Workload**
  - **Hadoop VS SQL**
  - **Which analytical database**
  - **Why InfiniDB**

Great Mobile Products That Sense the World

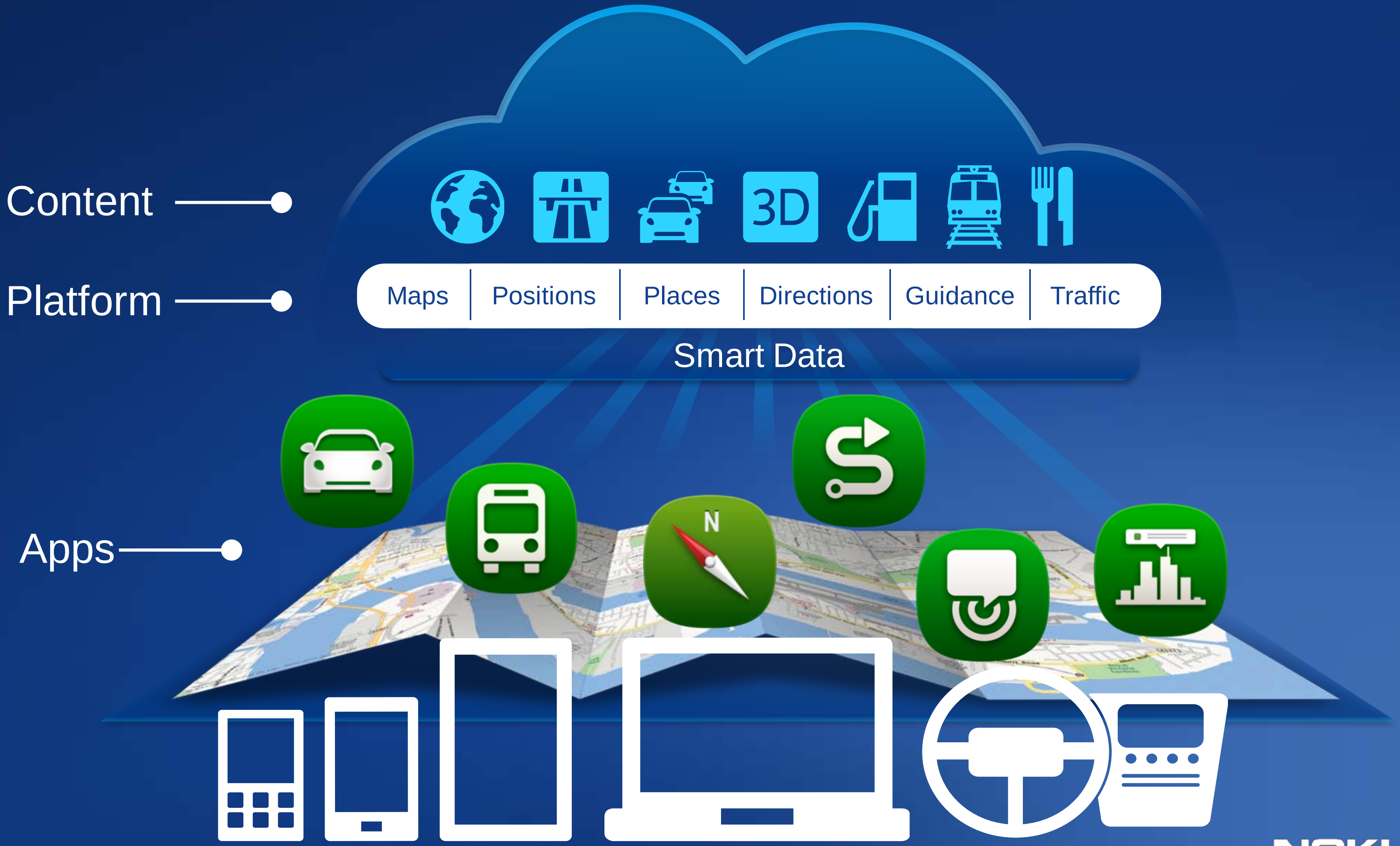
CREATE A LEADING “WHERE” PLATFORM

WIN IN SMART DEVICES

CONNECT THE NEXT BILLION

INVEST IN FUTURE DISRUPTIONS

One Platform, Enabling Contextually Rich Mobile Experiences



Big DATA ANALYTICS Platform @Nokia

Business Challenges

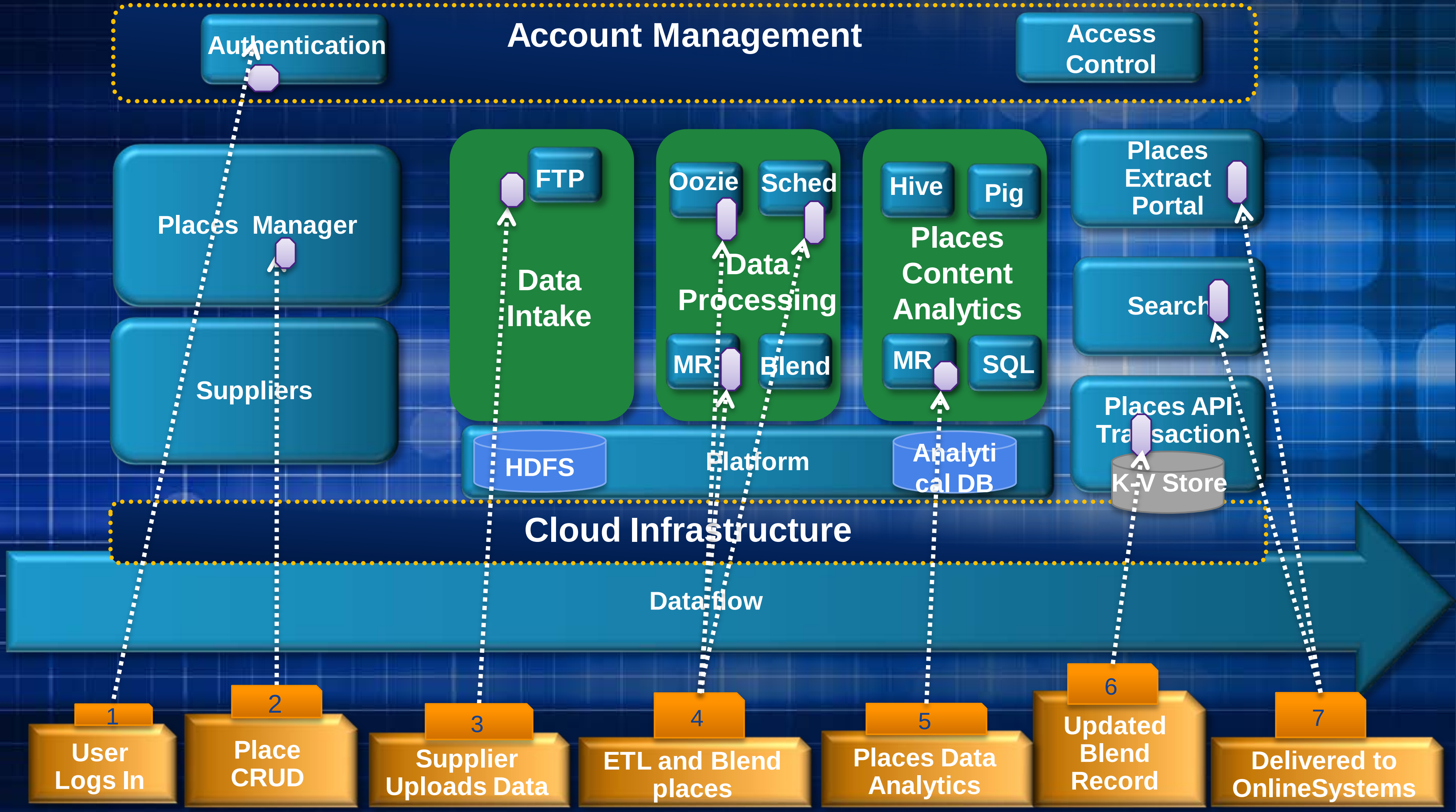
- Data silos, missing semantics
- Multiple sources - overlapping, conflicting
- Timely processing of large volumes of data
- Partial, insufficient, inaccurate, inconsistent.. data
- Security, privacy and other policies unknown

Central Analytics Platform created!

Statistics

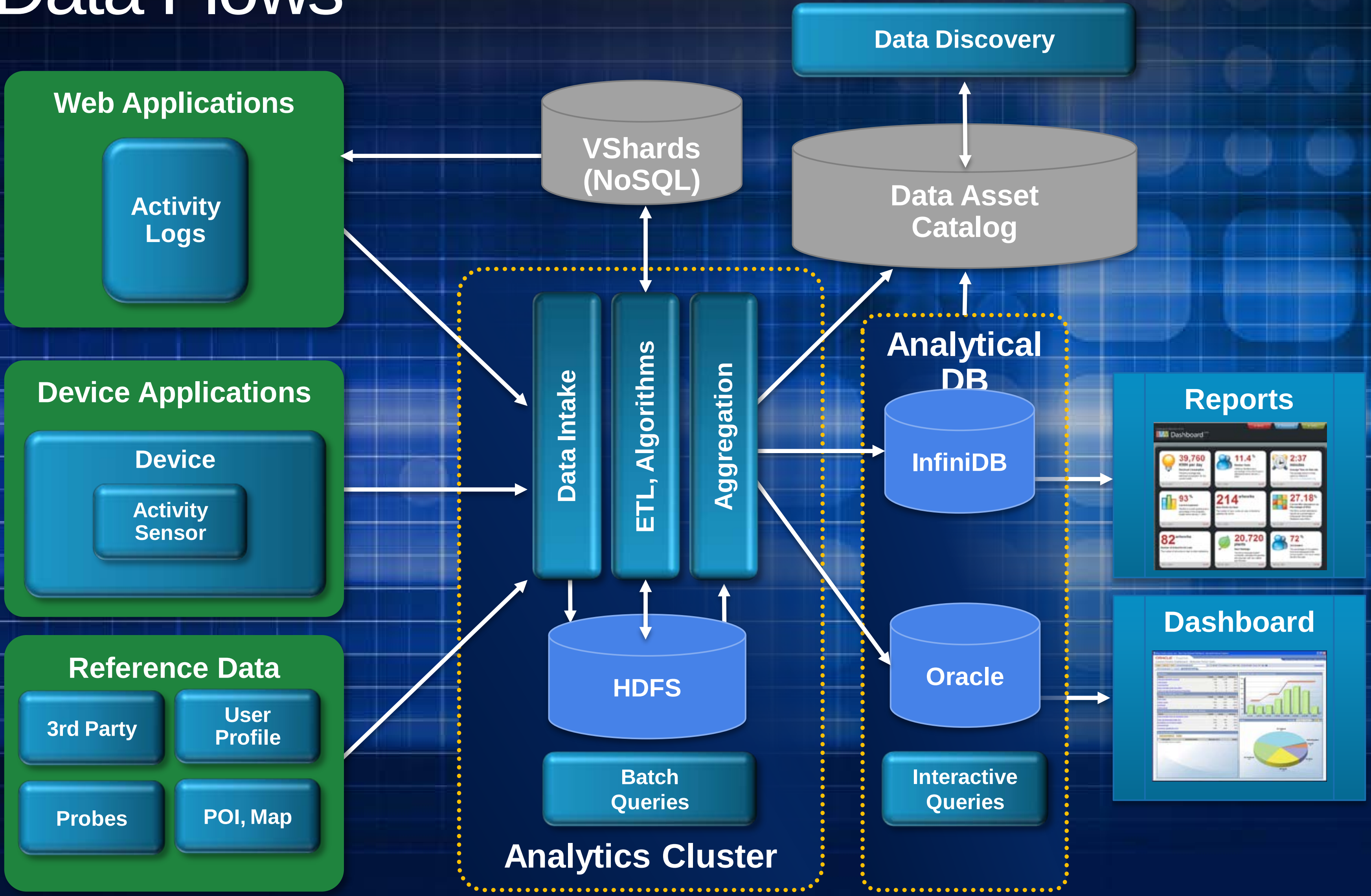
- 10’s PB of data all across Nokia
- Multi-tenant, multi-petabyte analytics cluster
- 10-20K+ jobs per day
- 600+ internal users
- 250M+ KV queries
- Over a terabyte flowing every day
- Multiple data centers around the world

Places Data Store (POI)- Use Case



NOKIA

Big Data Analytics Platform
Data Flows



Web Applications

Activity Logs

Device Applications

Device

Activity Sensor

Reference Data

3rd Party

User Profile

Probes

POI, Map

VShards (NoSQL)

Data Intake

ETL, Algorithms

Aggregation

HDFS

Batch Queries

Analytics Cluster

Data Discovery

Data Asset Catalog

Analytical DB

InfiniDB

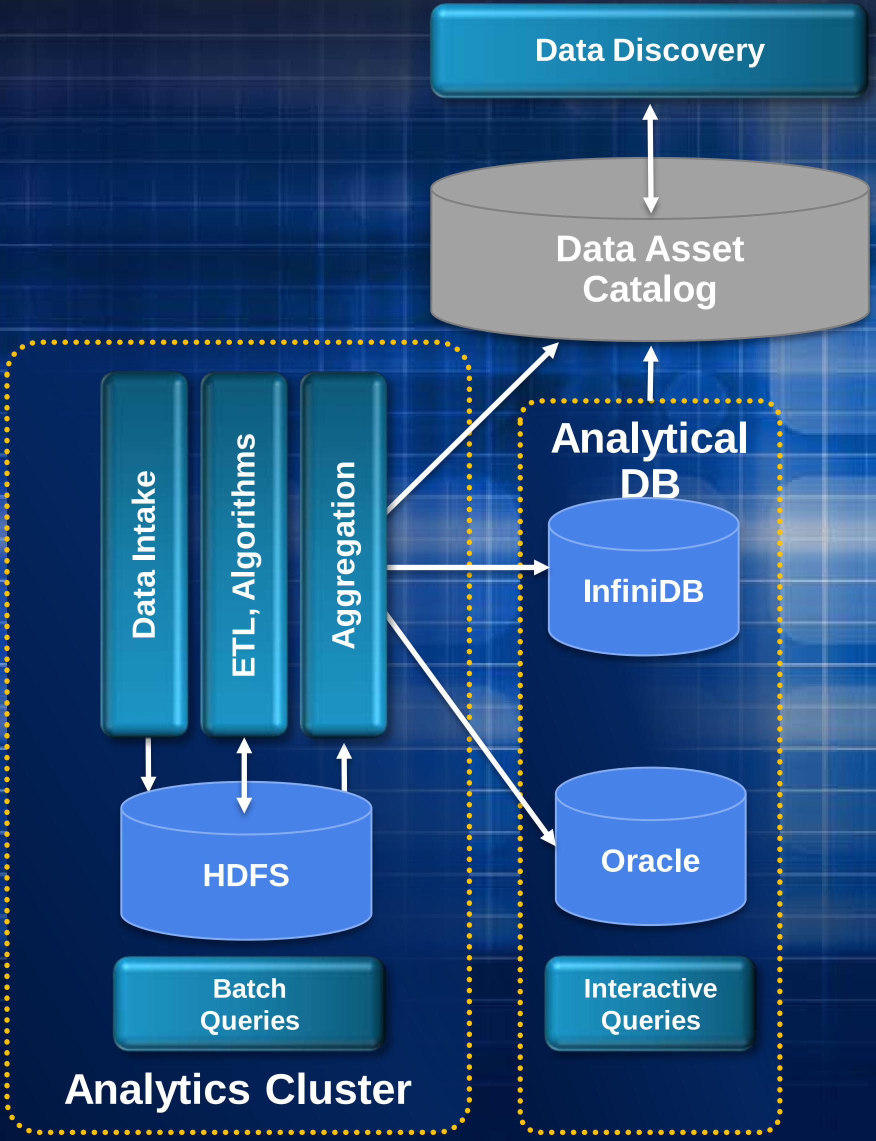
Oracle

Interactive Queries

Reports

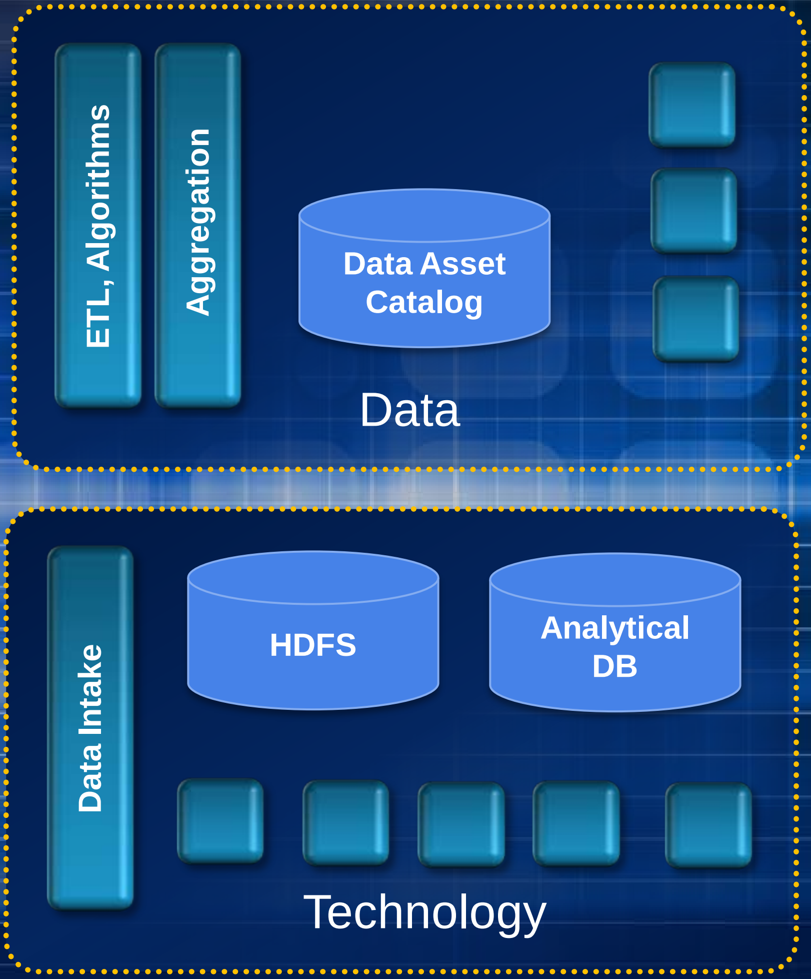
Dashboard

Big Data Analytics Platform
Data Flows

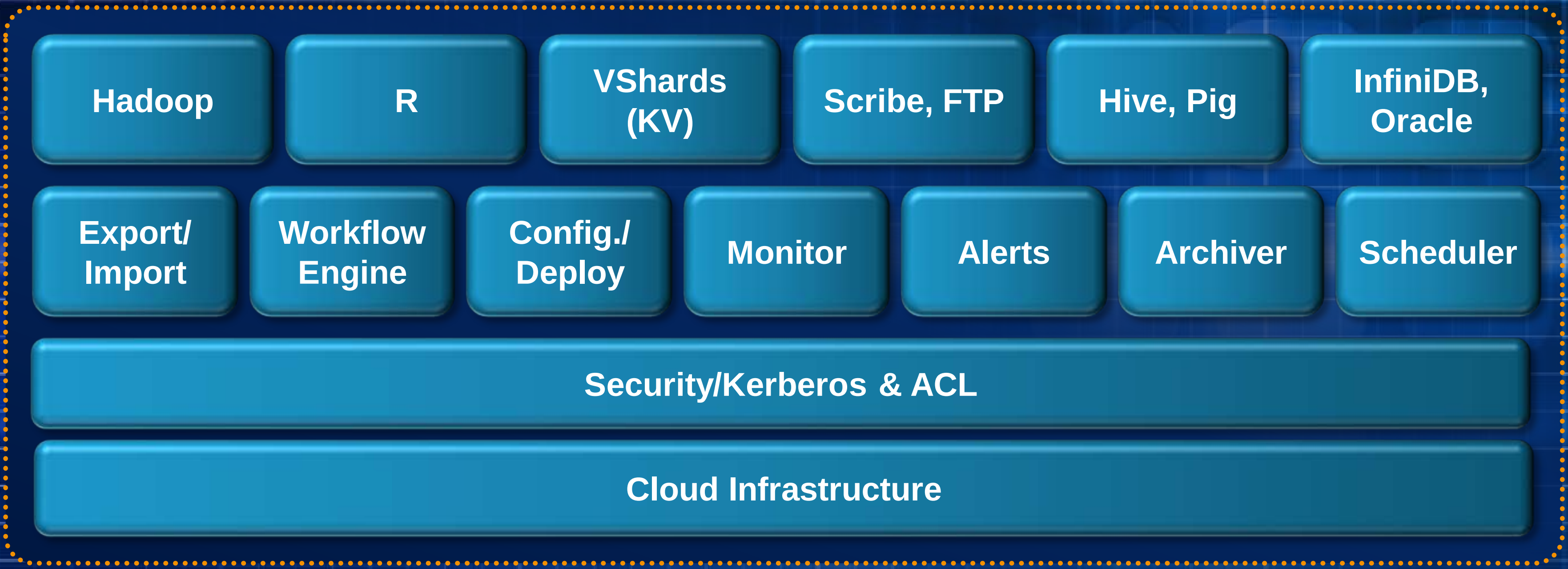


Big Data Analytics Platform

- **Logical Tiers**
  - **Technology Platform**
  - **Data Platform**
  - **End User Layer (not shown)**



Technology Platform



Hadoop

R

VShards (KV)

Scribe, FTP

Hive, Pig

InfiniDB, Oracle

Export/ Import

Workflow Engine

Config./ Deploy

Monitor

Alerts

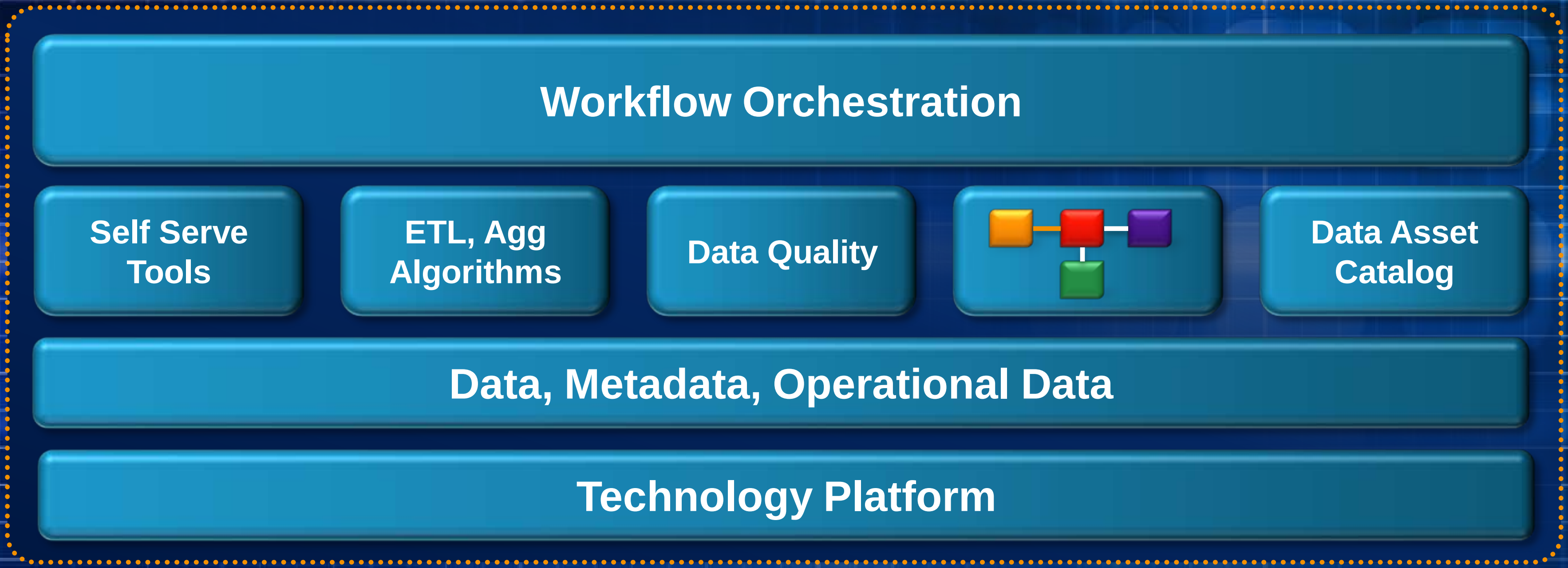
Archiver

Scheduler

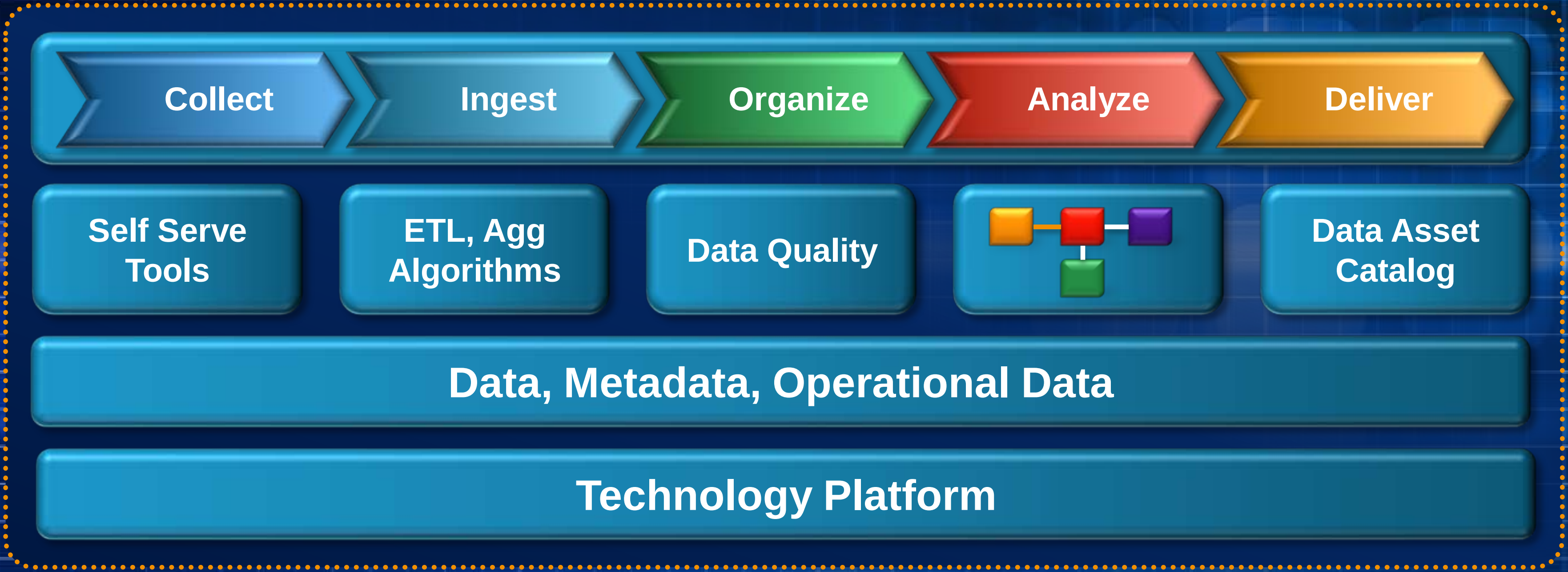
Security/Kerberos & ACL

Cloud Infrastructure

Data Platform



Data Platform – Analytics Lifecycle



Collect

Ingest

Organize

Analyze

Deliver

Self Serve Tools

ETL, Agg Algorithms

Data Quality

Data Asset Catalog

Data, Metadata, Operational Data

Technology Platform

Data Platform: Managing the Data Asset

- **Data Quality - garbage in , garbage out**
  - **Rules for validating, cleaning data, other heuristics**
  - **Trusting your insights**
  - **Process Quality**
  - **Light weight governance (semantics, integrity, privacy and quality)**

- **Data Asset Catalog – describe your data**
  - **Capture essential metadata and logical domain models for assets**
    - **physical model, logical model, policies, classifications**
    - **dependencies with other assets**
  - **Serves as a entry-point to data browsing and asset discovery**
  - **Insulates subject matter experts from physical details of data asset**

Big Data Challenges

- **At every level - capture, curate, storage, process, visualize..**
- **Hadoop or SQL ?**
  - **Performance of analytical database ?**
  - **Batch or Interactive analysis**
  - **Neither SQL nor MR fits all problems**
- **Data & Metadata Fragmentation**

Selecting the Right Tool for the Right Workload

Hadoop VS SQL/Analytical DB

SQL/Analytical DB

- Standard industry tools
- Interactive/Fast (secs)
- No coding, e.g. built-in functions
- Reasonable complex
- Discover the question

Hadoop/Hive/MR

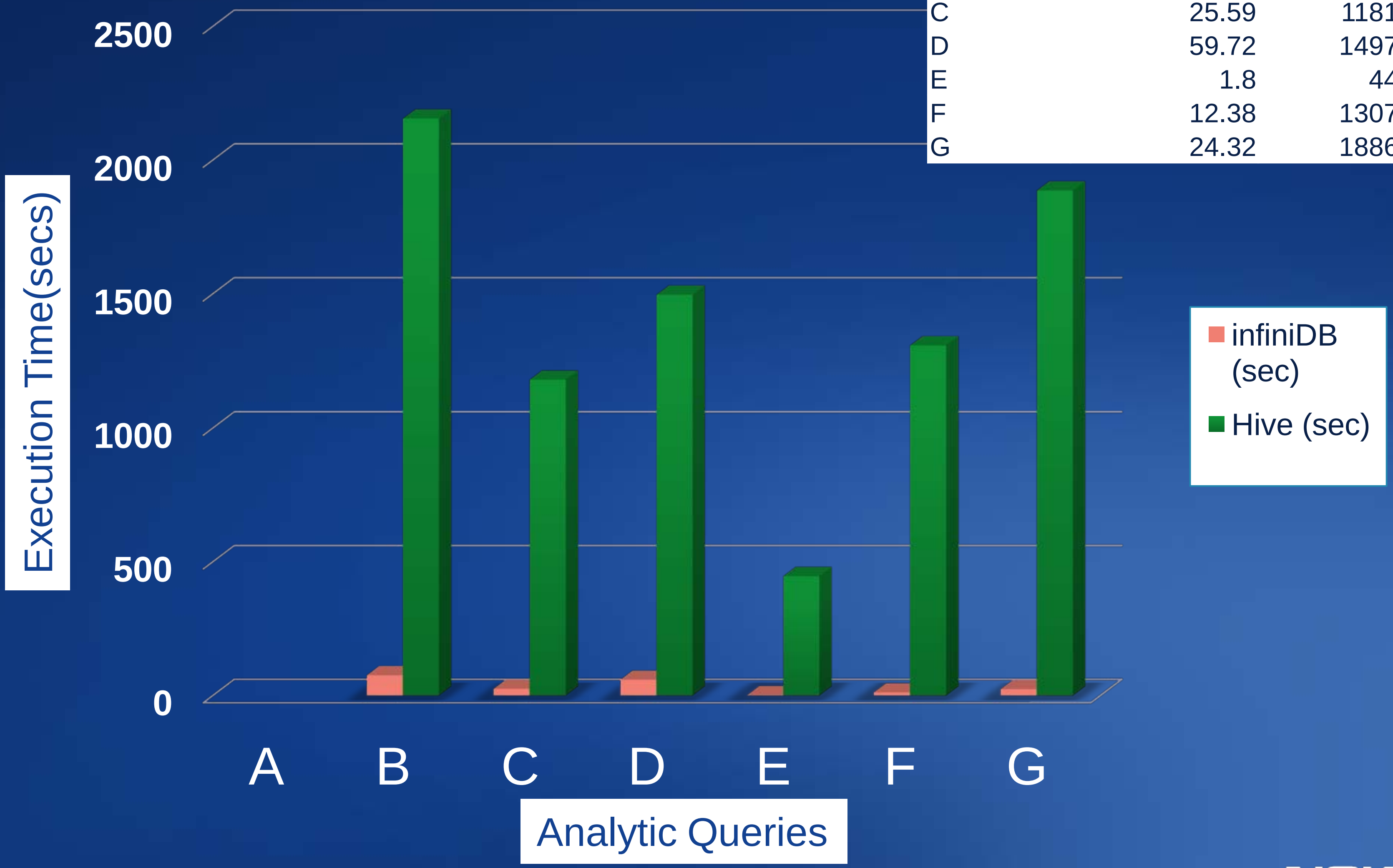
- ETL on steroids, Scale
- Batch/slow
- Bunch of coding, arbitrary complex
- Harvest & load into DW
- Discover the answer

Why InfiniDB ?

- Cloud deployment model
- Column oriented, MPP, clean architecture
- Horizontal and vertical partitioning, clever pruning
- No indexes
- Efficient joins
- Impressive benchmarks
- Stream based MR like processing
- Works with BI tools (standard JDBC driver)

InfiniDB vs Hive Performance

Query	InfiniDB (sec)	Hive (sec)
A		
B	76.32	2155.92
C	25.59	1181.48
D	59.72	1497.22
E	1.8	446.5
F	12.38	1307.38
G	24.32	1886.81

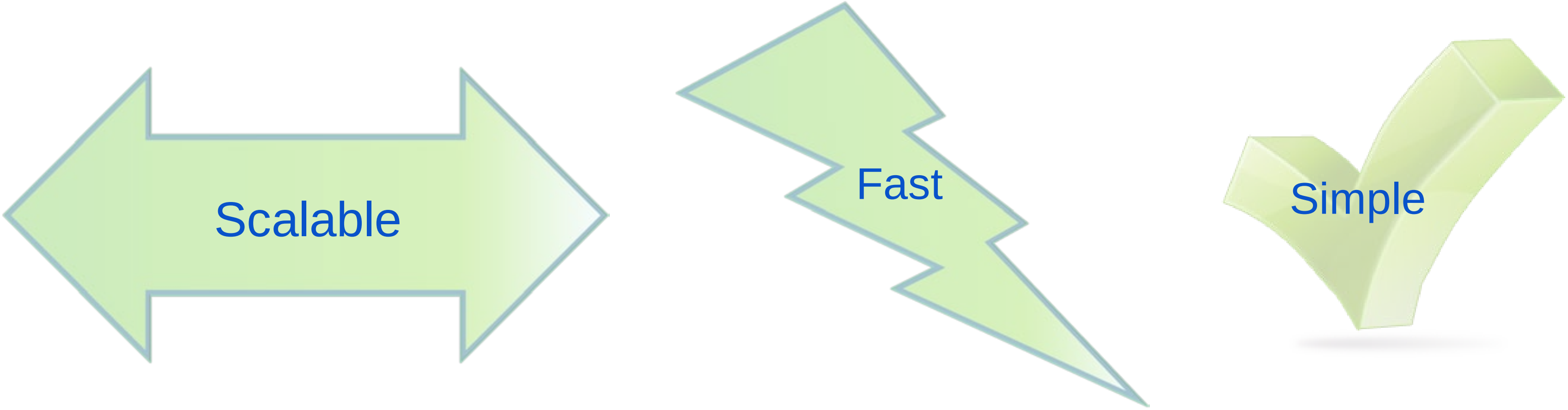


InfiniDB Under the Hood

What is InfiniDB?

InfiniDB®

Scalable. Fast. Simple.



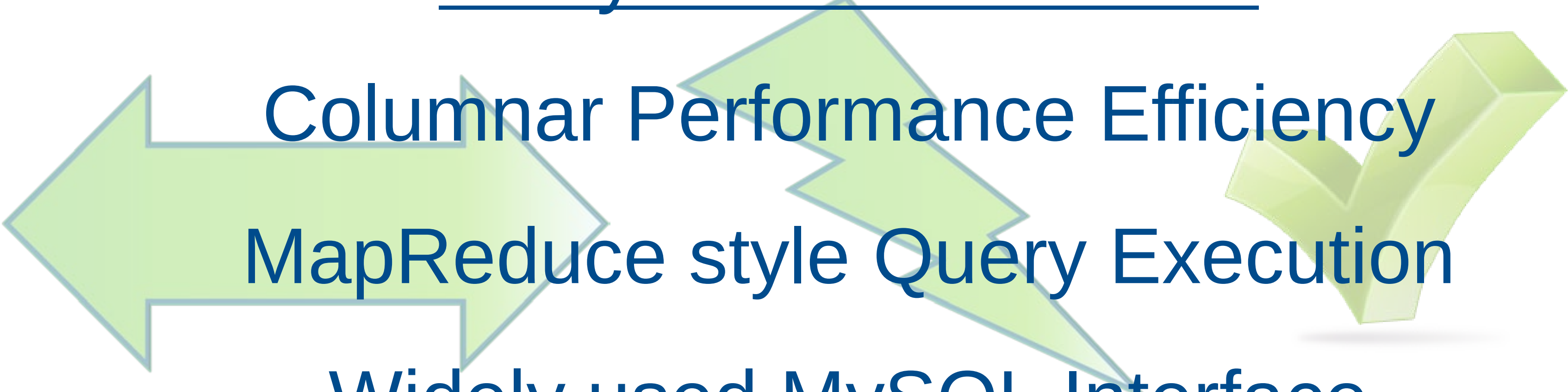
Scalable

Fast

Simple

Analytics Data Platform Foundation

InfiniDB®



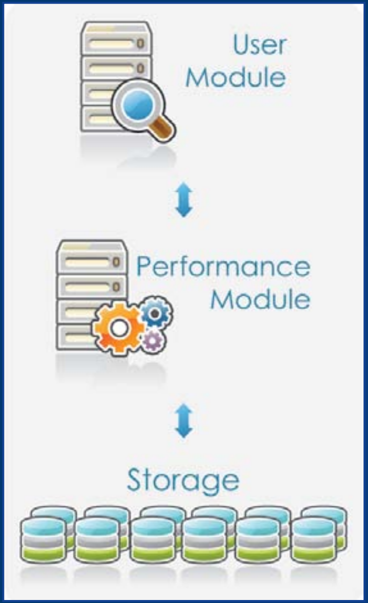
Analytics Data Platform

Columnar Performance Efficiency

MapReduce style Query Execution

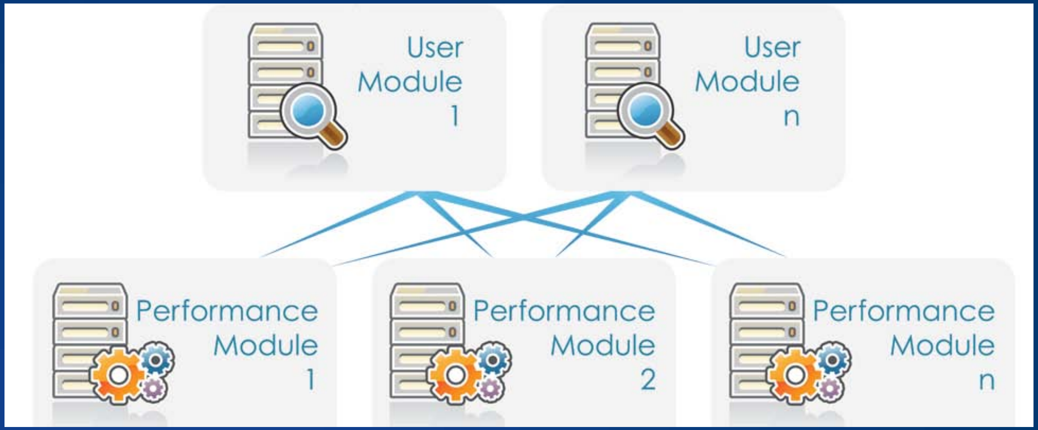
Widely used MySQL Interface

InfiniDB Building Blocks



Single Server

or …

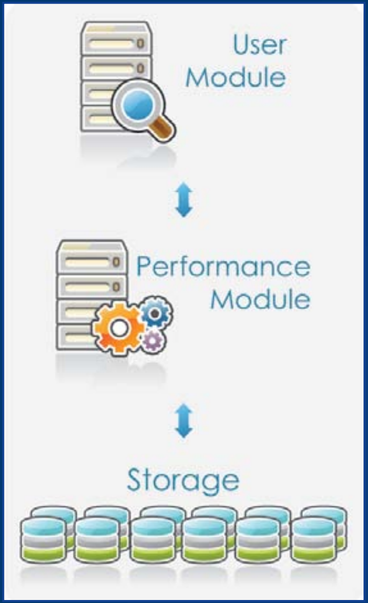


Purpose built for big data analytics.

- User Module (UM)

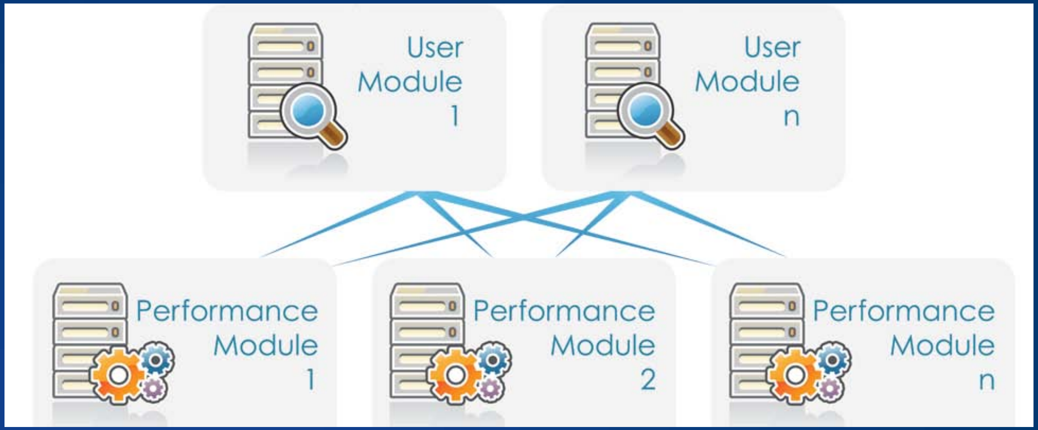
- Performance Module (PM)

InfiniDB Building Blocks



Single Server

or …



Purpose built for big data analytics.

- User Module (UM)

  Understands SQL

- Performance Module (PM)

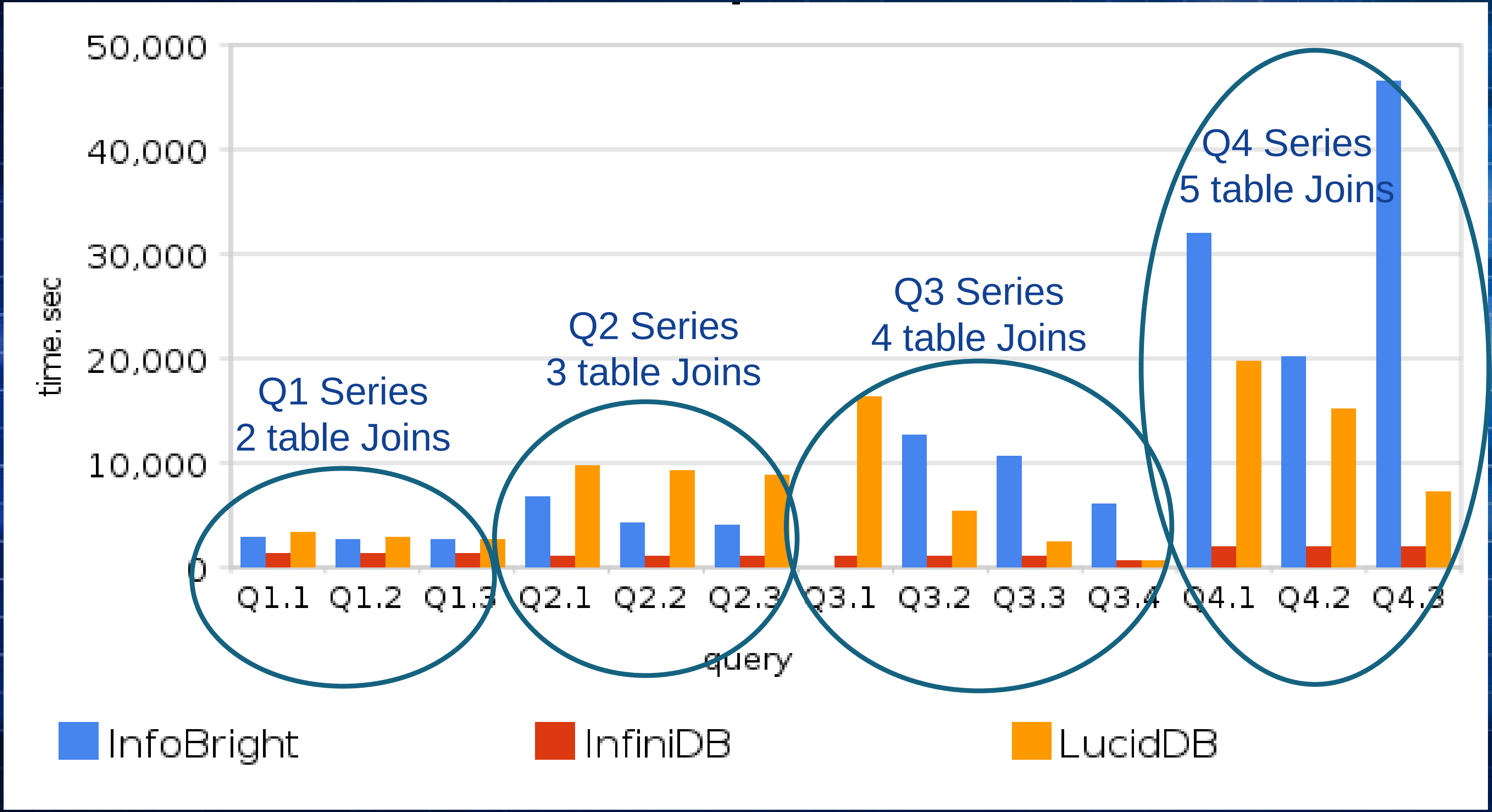
  Operates on data blocks

calpont
ACCELERATING DATA INSIGHTS

InfiniDB M/R Style Distribution of Work “Map-Reduce Inside”

	InfiniDB DoW	Hadoop M/R
Scalability	Linear	Linear
N-squared Problem	Avoided	Avoided
Latency	Low	Medium-High
Intermediate Results Handling	Stream-based	File-based
Report Language	SQL	Erlang M/R, Hive, Pig
Tuning	Automatic	Manual
Real-Time Analytics	Real-time access to granular data	Access to pre-defined aggregates
Ad-Hoc	Full Ad-Hoc performance	None
Data Storage	Structured	Unstructured

Independent InfiniDB Benchmark



Q1 Series
2 table Joins

Q2 Series
3 table Joins

Q3 Series
4 table Joins

Q4 Series
5 table Joins

Takeaways

- Hadoop is good but….
- Pay attention to data quality
- Hadoop or SQL
- Describe your data

THANK YOU

Yekesa Kosuru
Distinguished Architect, Nokia
yekesa.kosuru@nokia.com
www.nokia.com
@Nokia

Jim Tommaney
CTO, Calpont
jtomanney@calpont.com
www.calpont.com
@Calpont, @InfiniDB

calpont
ACCELERATING DATA INSIGHTS

NOKIA