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

Nokia Internal Use Only

Great Mobile Products That Sense the World

CREATE A LEADING “WHERE” PLATFORM

WIN IN SMART DEVICES

CONNECT THE NEXT BILLION

INVEST IN FUTURE DISRUPTIONS

Nokia Internal Use Only
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

1. User Logs In
2. Place CRUD
3. Supplier Uploads Data
4. ETL and Blend places
5. Places Data Analytics
6. Updated Blend Record
7. Delivered to Online Systems
Big Data Analytics Platform
Data Flows

Web Applications
- Activity Logs

Device Applications
- Device
- Activity Sensor

Reference Data
- 3rd Party
- User Profile
- Probes
- POI, Map

Data Flows:
- Data Intake
- ETL, Algorithms
- Aggregation
- HDFS
- Batch Queries

Analytics Cluster:
- VShards (NoSQL)
- Oracle
- InfiniDB

Data Asset Catalog
- Interactive Queries
- Interactive Queries

Data Discovery
- Reports

VShards

Dashboard

Analytics DB

Web Applications

Device Applications

Reference Data
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

Workflow Orchestration

Self Serve Tools
ETL, Agg Algorithms
Data Quality
Data Asset Catalog

Data, Metadata, Operational Data

Technology Platform
Data Platform – Analytics Lifecycle

Collect
Self Serve Tools

Ingest
ETL, Agg Algorithms

Organize
Data Quality

Analyze
Data Asset Catalog

Deliver

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

<table>
<thead>
<tr>
<th>Query</th>
<th>InfiniDB (sec)</th>
<th>Hive (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>76.32</td>
<td>2155.92</td>
</tr>
<tr>
<td>B</td>
<td>25.59</td>
<td>1181.48</td>
</tr>
<tr>
<td>C</td>
<td>59.72</td>
<td>1497.22</td>
</tr>
<tr>
<td>D</td>
<td>1.8</td>
<td>446.5</td>
</tr>
<tr>
<td>E</td>
<td>12.38</td>
<td>1307.38</td>
</tr>
<tr>
<td>F</td>
<td>24.32</td>
<td>1886.81</td>
</tr>
</tbody>
</table>

Analytic Queries
InfiniDB Under the Hood
What is InfiniDB?

InfiniDB®

Analytics Data Platform Foundation

InfiniDB

Analytics Data Platform

Columnar Performance Efficiency

MapReduce style Query Execution

Widely used MySQL Interface
InfiniDB Building Blocks

Purpose built for big data analytics.

- User Module (UM)
- Performance Module (PM)
InfiniDB Building Blocks

Purpose built for big data analytics.

- User Module (UM)
  - Understands SQL
- Performance Module (PM)
  - Operates on data blocks
### InfiniDB M/R Style Distribution of Work “Map-Reduce Inside”

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