Using Big Data to Bridge the Virtual & Physical Worlds

Ishita Gupta - 2019DMF05
Sri Rajitha - 2019DMB09
Yogesh Karnam - 2019DMB04
Rohith Krishna - 2019DMB07

About the company

- Telecommunication company that collects, stores and analyzes user data.
- Builds maps with predictive traffic and layered elevation models (for their own map application and related services.)
- Sources information about points of interest around the world.
- Collaboration with Microsoft has accelerated digital transformation for enterprises, for instance,
    - Connected smart tools,
    - Robots in industrial environment,
    - Automated machines to improve efficiency etc.
- Performs complex analyses to understand the quality of phones, etc.

Nokia deals with

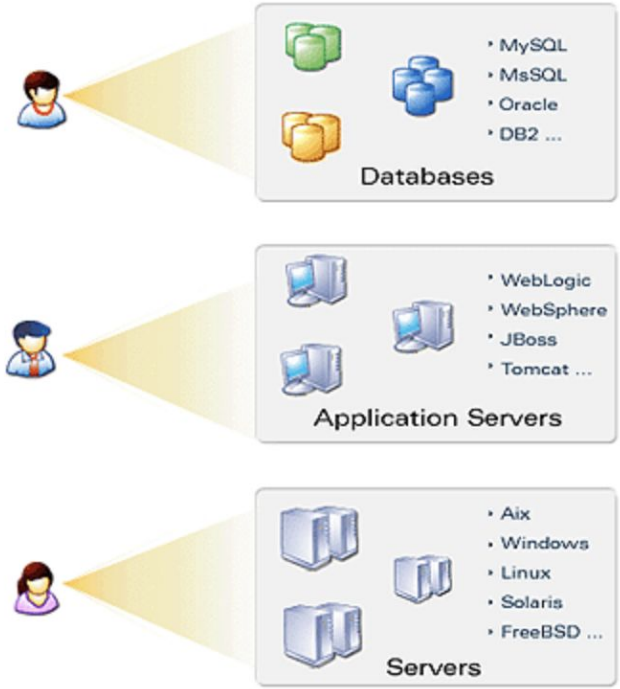
- More than 500 terabytes of unstructured data
- Close to 100 terabytes of structured data.

Business Problems

- Nokia deals with
  - More than 500 terabytes of unstructured data
  - Close to 100 terabytes of structured data.
- They initially had numerous application silos. Associated issues:
  - Different versions/databases
  - Integrating these databases could be useful for cross-referencing.
  - Single holistic repository of data from which relevant insights could be derived.
- Computational cost associated with processing this vast database.
- PB-level data with relational database:
  - Costly
  - Limits the data types / data volume used.

Information Silo

- An information silo is created when departments or groups within an organization choose not to share information or allow for knowledge to be exchanged through information systems with other groups of individuals in the same organization.
- As groups work separately and continue to restrict shared access to information and systems it becomes more difficult to create a consensus on priorities for the entire company.



Business Solutions - Components of Hadoop

- Use of Cloudera's Distribution containing Apache Hadoop (CDH).
  - Bundles existing open-source Apache Hadoop frameworks.
- HDFS
  - System that stores all the structured and unstructured datasets
  - Allows processing of the stored data at a petabyte scale.
- HBASE - scalable, distributed database that supports structured data storage for large tables.
- Scribe - a server for aggregating log data streamed in real-time from many servers.
  - Designed to be scalable, extensible without client-side modification
  - Robust to failure of the network or any specific machine.
- Sqoop - efficiently transferring bulk data between Hadoop and structured datastores such as relational databases.

Business Solutions - Data Warehouse

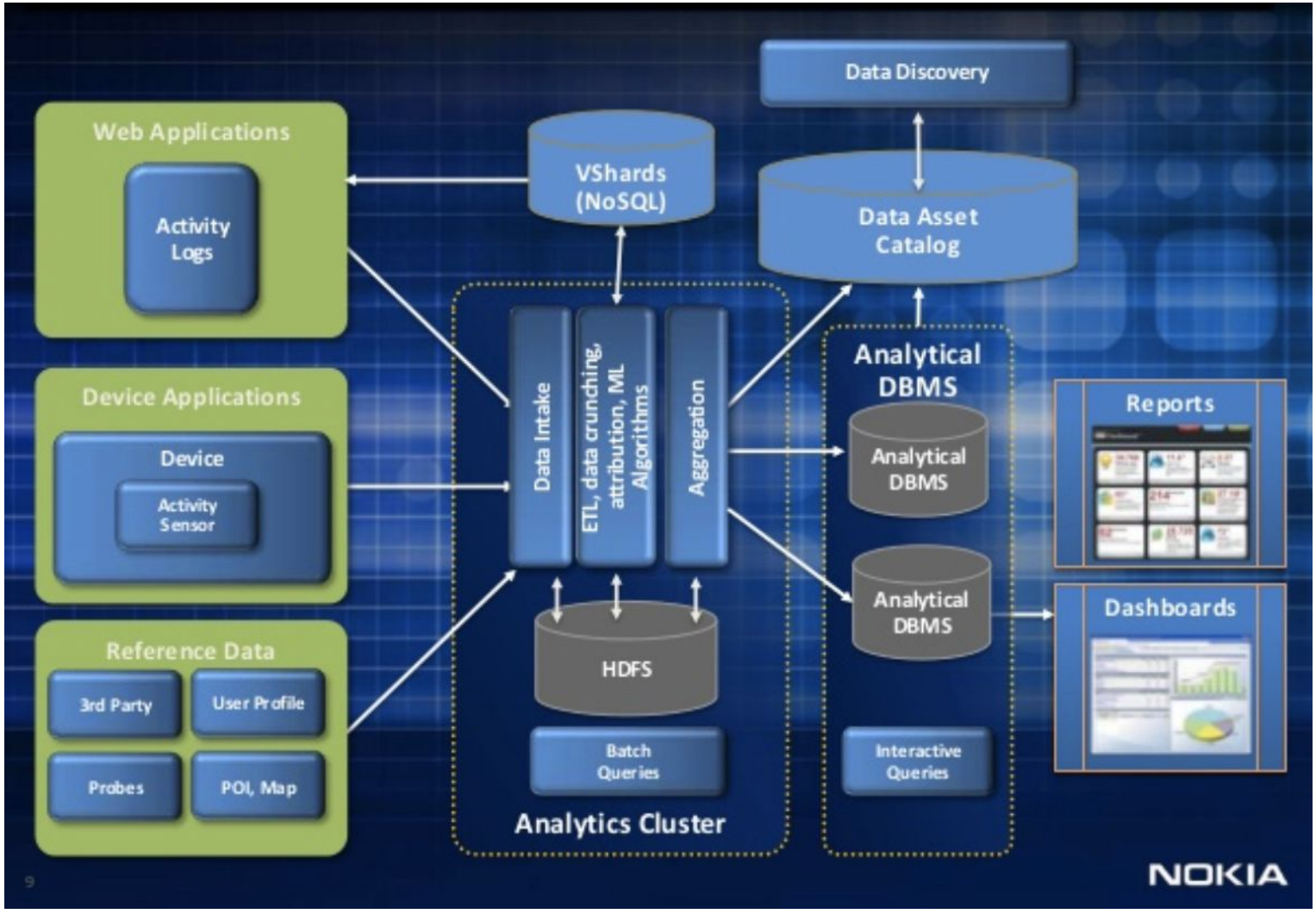
- A data warehouse -
  - Constructed by integrating data from multiple heterogeneous sources
  - Supports analytical reporting, structured and/or ad hoc queries, and decision making.
- Nokia’s data warehouses and marts
  - Continuously stream multi-structured data into a multi-tenant Hadoop environment,
  - Allows the company’s 60,000+ employees to access the data.
- Teradata - Data warehousing solution
  - Performs data integration efficiently by using parallel processing.
  - Supports SQL queries
  - delivers real-time, intelligent answers using Pervasive Data Intelligence
  - Teradata Vantage integrates with popular third-party tools & analytic languages.

Business Solutions - Servers

- Dell PowerEdge R510 and R710 servers:
  - Designed to meet the needs of many IT environments
  - Advanced systems management capabilities,
  - A compact chassis, high-availability and redundancy features, and large amounts of internal storage capacity.

Business Solutions

- 100 terabytes (TB) of structured data on Teradata and petabytes (PB) of multi-structured data on the Hadoop Distributed File System (HDFS),
- Separate warehouses for structured and unstructured data - maintained using Teradata.
- HDFS allows storing all the semi/multi structured and unstructured data and offers processing of the stored data at petabyte scale.
- Scribe processes - moves data efficiently between components. Eg. servers in Singapore to a Hadoop cluster in the UK data center.
- Sqoop - used to move data from HDFS to Oracle and/or Teradata.
- HBase - used to serves data out of Hadoop. Eg. Unstructured data such as log files, billions of rows of call records.


Data Discovery
Web Applications
Activity Logs
VShards (NoSQL)
Data Asset Catalog
Device Applications
Device
Activity Sensor
Data Intake
ETL, data crunching, attribution, ML Algorithms
Aggregation
Analytical DBMS
Analytical DBMS
Reports
Analytical DBMS
Dashboards
Reference Data
3rd Party
User Profile
Probes
POI, Map
HDFS
Batch Queries
Interactive Queries
Analytics Cluster
NOKIA

Benefits

- Cost per terabyte is 10x cheaper than traditional relational warehouse system.
- Reliable, cost-effective data-storage solution.
- High performance parallel processing of multi-structured data at petabyte scale.
- Accepts variety of data and derives value from diverse data.
- Fault tolerance in HDFS
    - Fail-safe in case of damage to name node.
    - When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use in the secondary name node.

Limitations

- Unstructured data must be re-formatted to fit into a relational schema before it can be loaded into the system.
  - Requires extra data processing step.
  - Slows ingestion, creates latency and eliminates potentially important data.
- Difficult to deploy in production.
- Hadoop is poor on query performance and data management.
- MapReduce requires a lot of time to perform tasks.
  - Increases latency.
  - Reduces processing speed.

Conclusions

- Nokia currently uses Hadoop to creating 3D digital maps that incorporate traffic models that understand speed categories, recent speeds on roads, historical traffic models, elevation, ongoing events, video streams of the world, and more.
- Cloudera’s platform is essential for Nokia’s engineering, strategic support and training services teams to derive marketable business insights using Big Data.