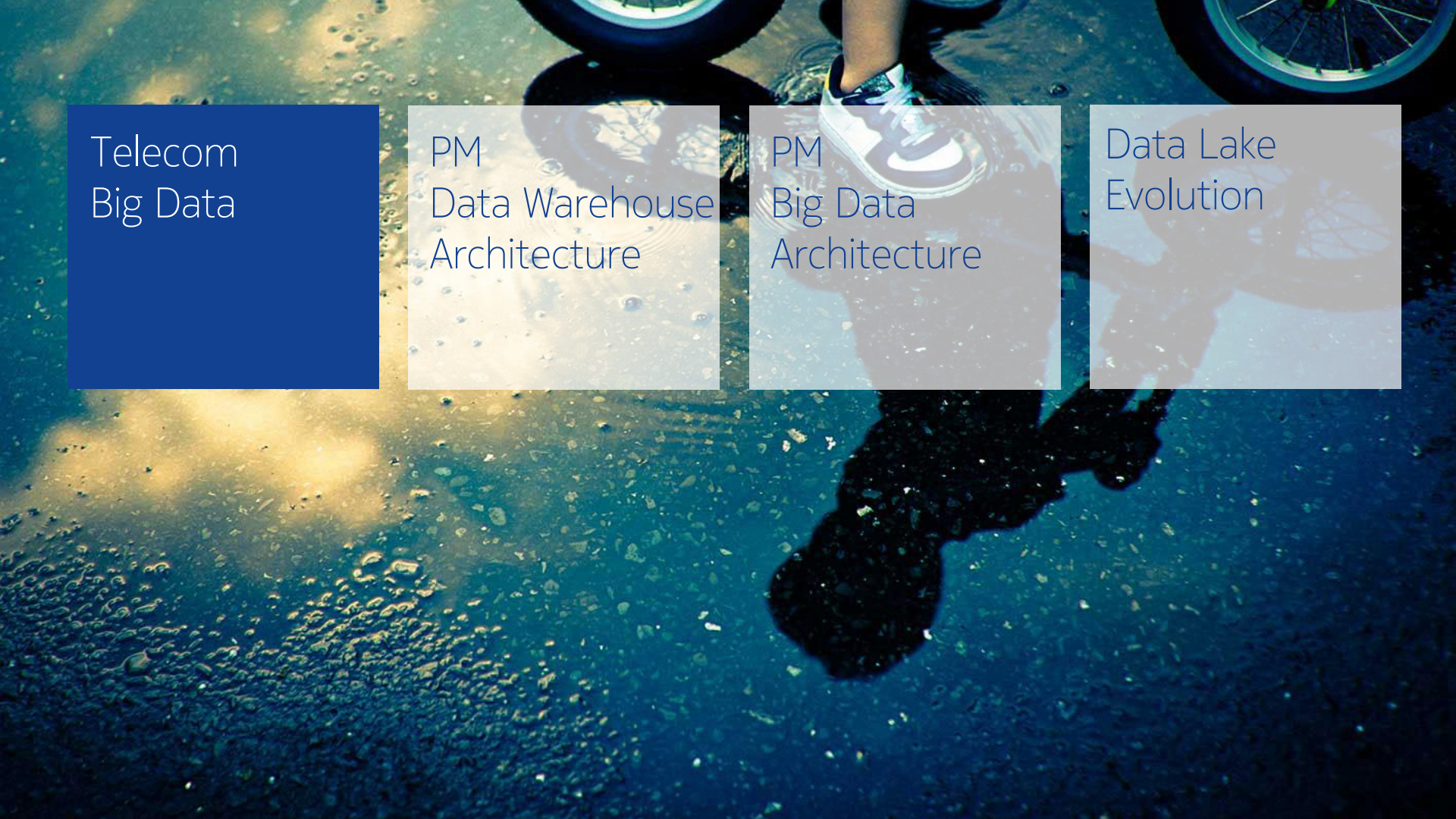
From Data Warehouse to Data Lake

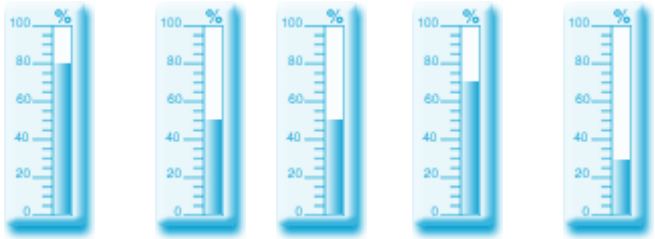
——Telecom Big Data Architecture evolution

石涛声
NOKIA成都研发中心
2016.10.15


Telecom
Big Data
PM
Data Warehouse
Architecture
PM
Big Data
Architecture
Data Lake
Evolution

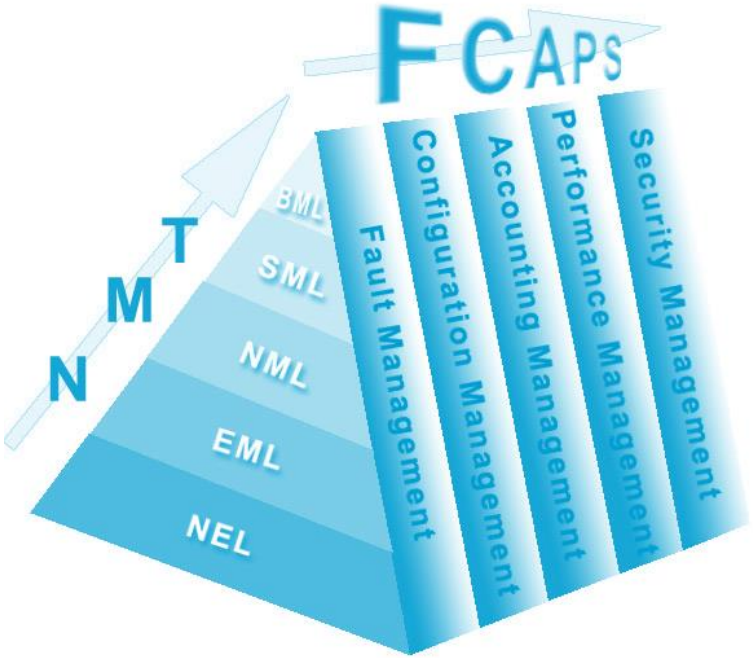
The TMN/FCAPS Model

Current mobile network architecture



F C A P S

Fault Management
Configuration Management
Accounting Management
Performance Management
Security Management



NOKIA

Enable technology
Technique Insight

Most structured data

Variety

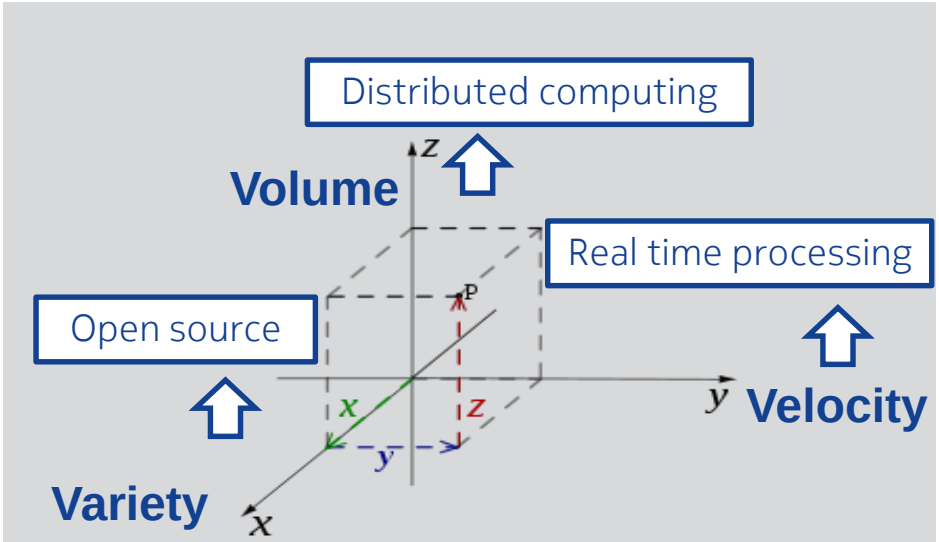
Massive network data

Volume

Every second in real-time

Velocity

Real-time measurement, analytics, decision and action

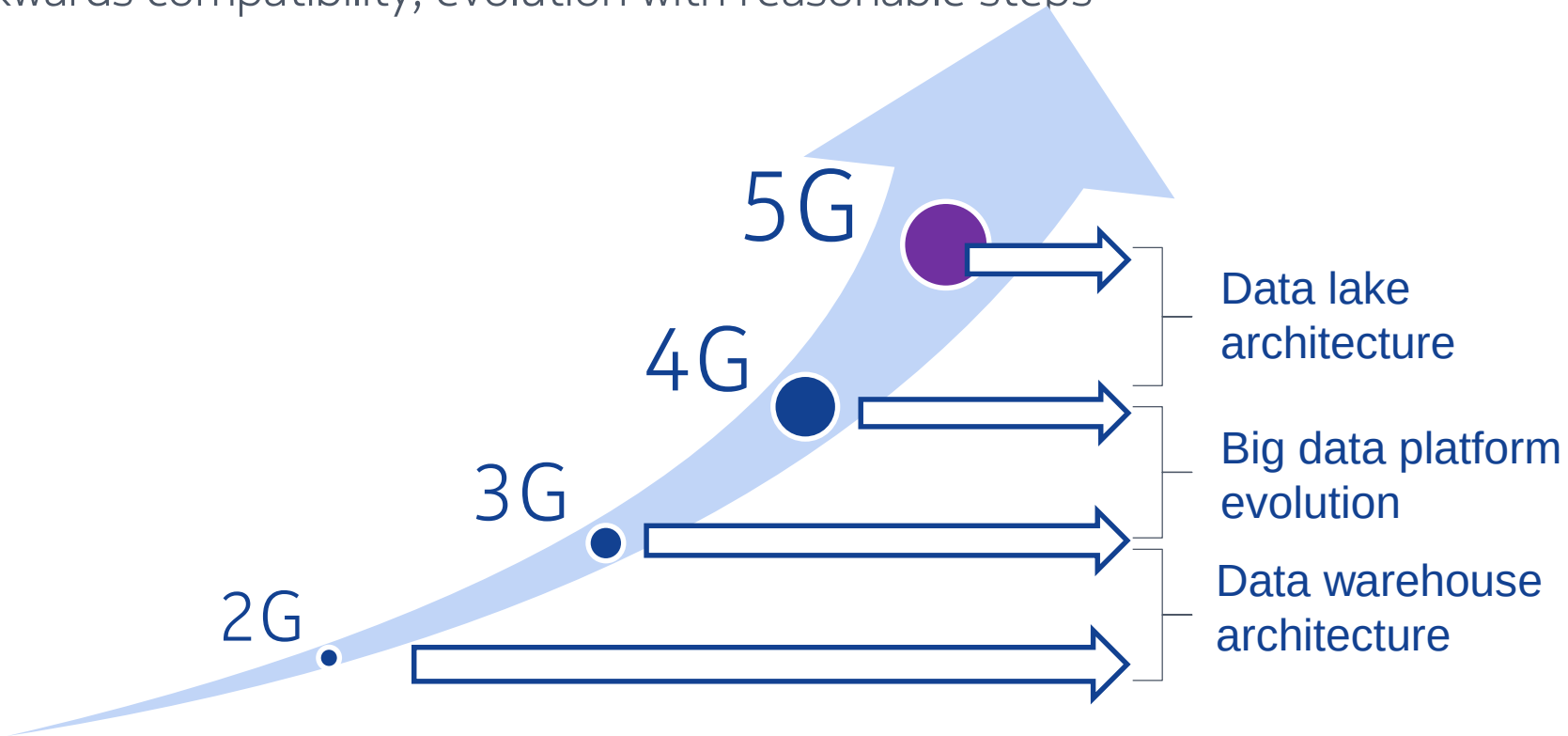


Three dimension of BigData

- The streaming paradigm to deal with the velocity of the data
- Distributed computing to deal with the volume of the data
- open source for the variety

Stepwise approach

Backwards compatibility, evolution with reasonable steps

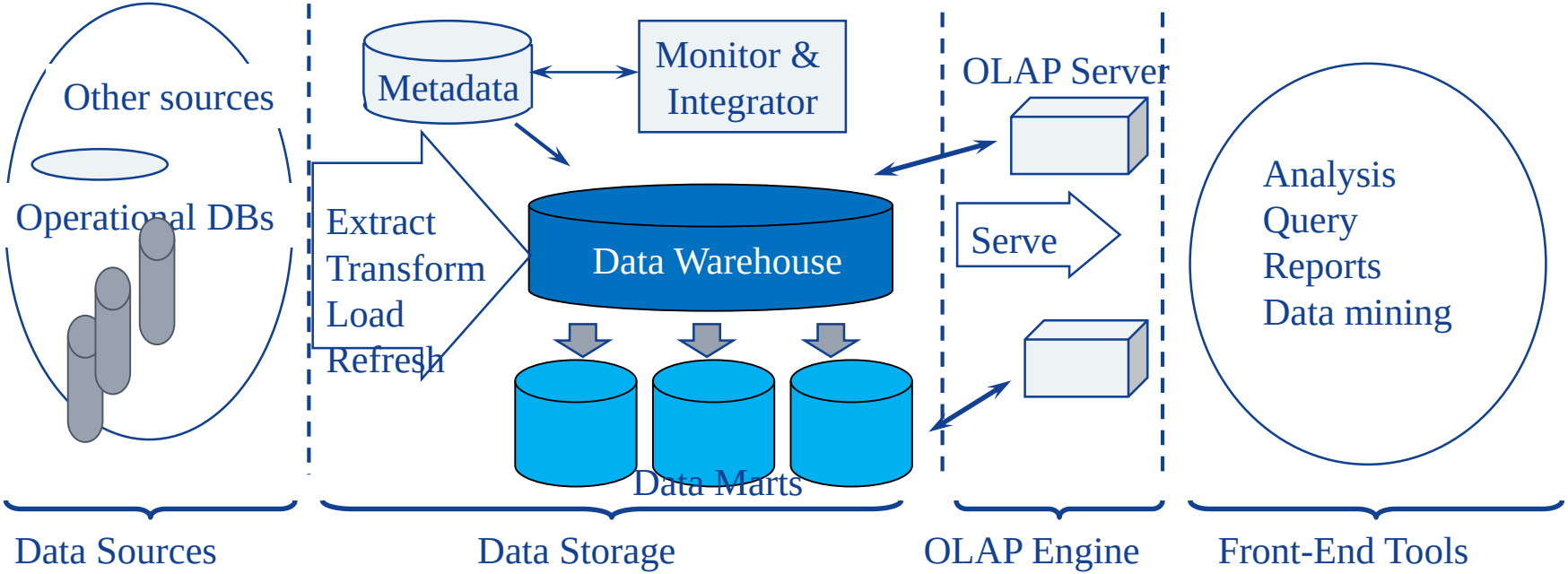



Telecom
Big Data
PM
Data Warehouse
Architecture
PM
Big Data
Architecture
Data Lake
Evolution

What is Data Warehouse?

Typical data warehousing architecture : A Multi-Tiered Architecture

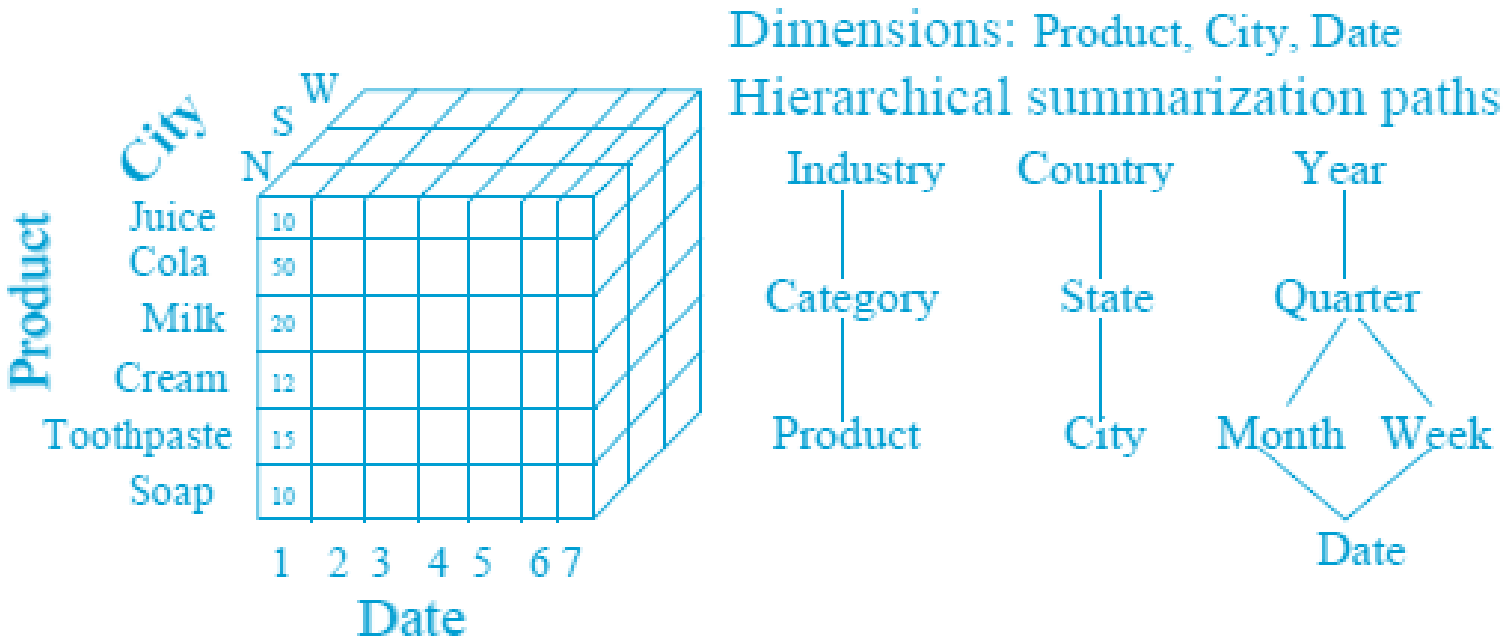
"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process." —W. H. Inmon



Multidimensional Data

Data cube

A dimension is a schema object that defines hierarchical relationships between columns or column sets.



Typical OLAP Operations

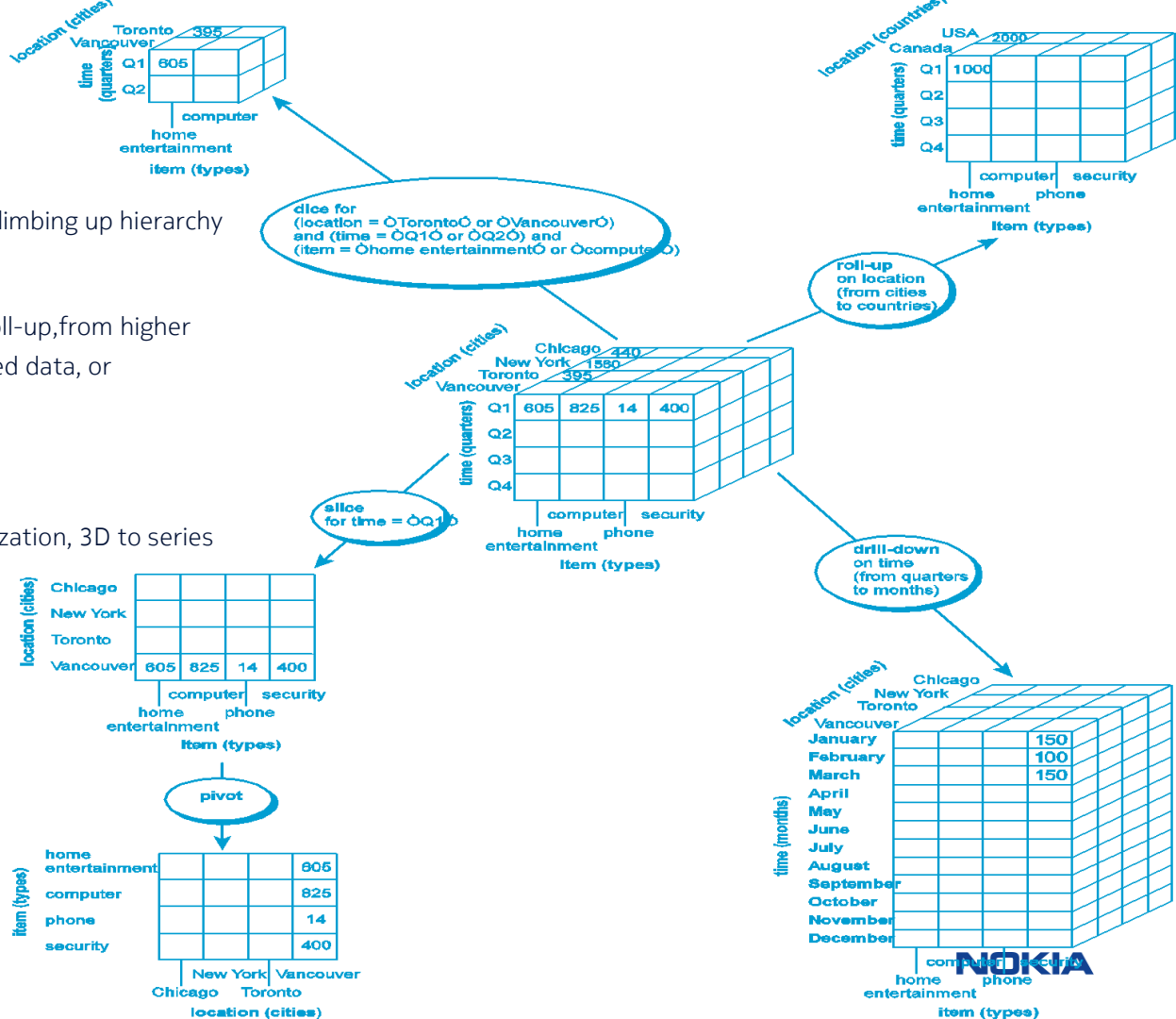
Typical OLAP Operations

Roll up (drill-up): summarize data by climbing up hierarchy or by dimension reduction

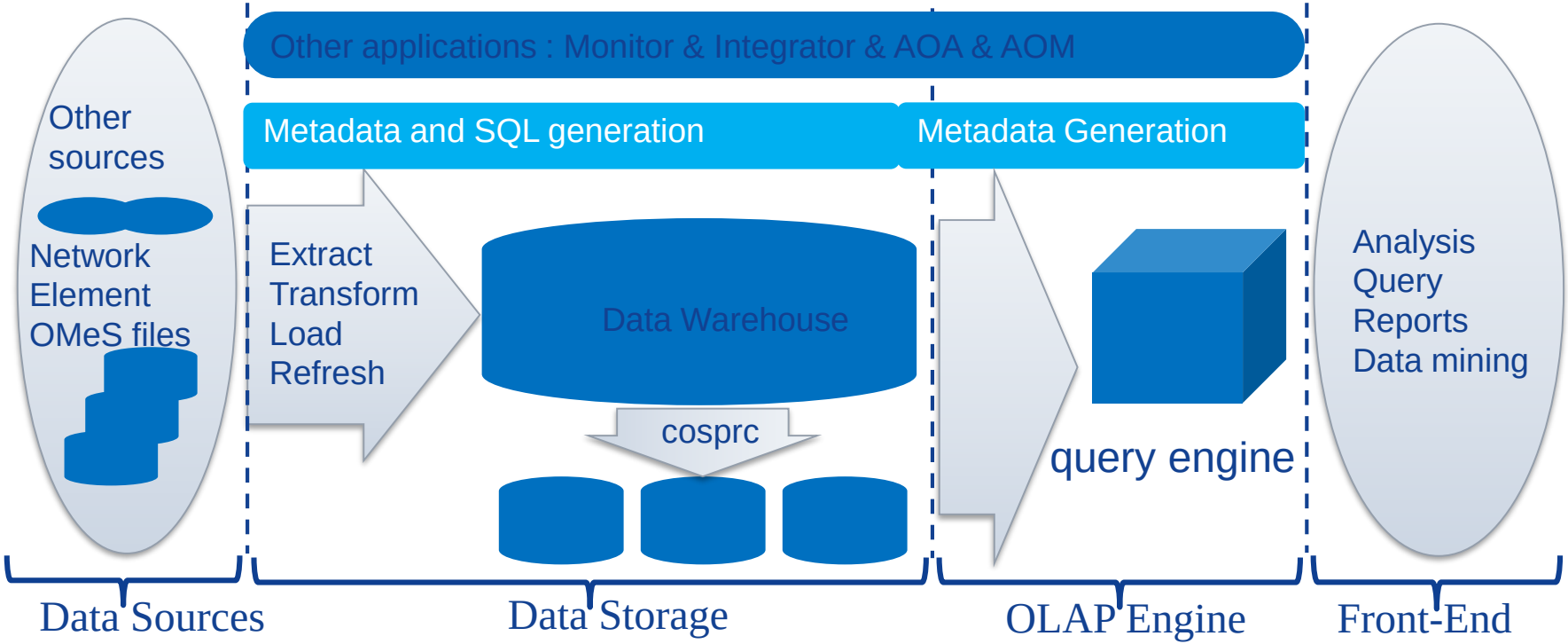
Drill down (roll down): reverse of roll-up,from higher level summary to lower level summary or detailed data, or introducing new dimensions

Slice and dice: project and select

Pivot (rotate): reorient the cube, visualization, 3D to series of 2D planes



Reference architecture
Meta Data Driven Architecture Design



NOKIA

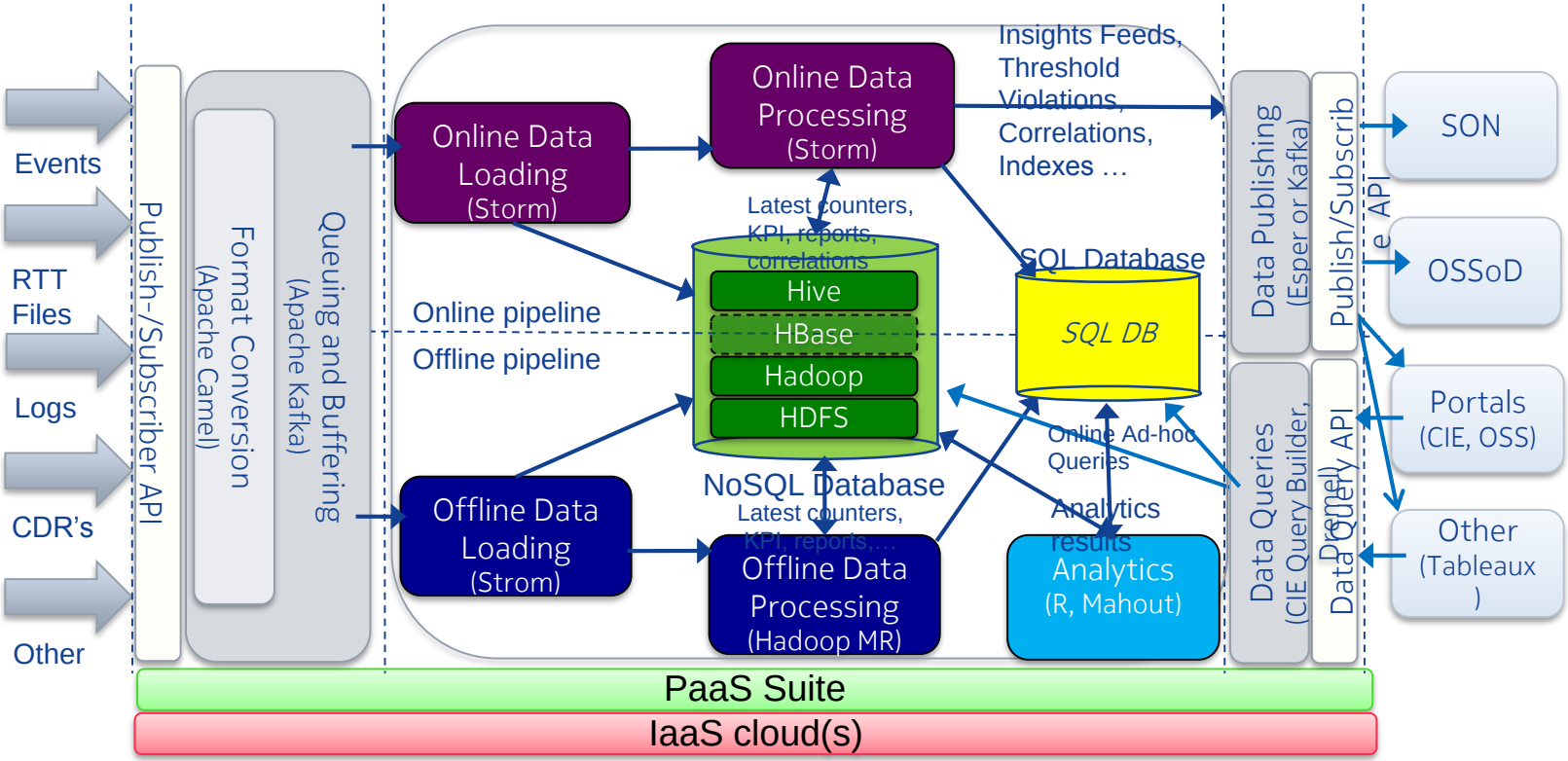
Pain Point of current data warehouse architecture
From Oracle RAC to Big Data stack

- Data warehouse is currently based on **Oracle relational database**
- DB scalability is handled via **Oracle RAC solution**, whenever **vertical scalability** is no longer possible (due to HW limitations)
- **Oracle RAC is not linear scalable** and introduces several challenges wrt existing deployments (e.g., administration overhead, Customer CoDe, …)
- Storage solution for proper **tuned I/O sub-system is quite expensive** and storing raw/hourly data for extended periods is also **cost prohibitive**

NOKIA

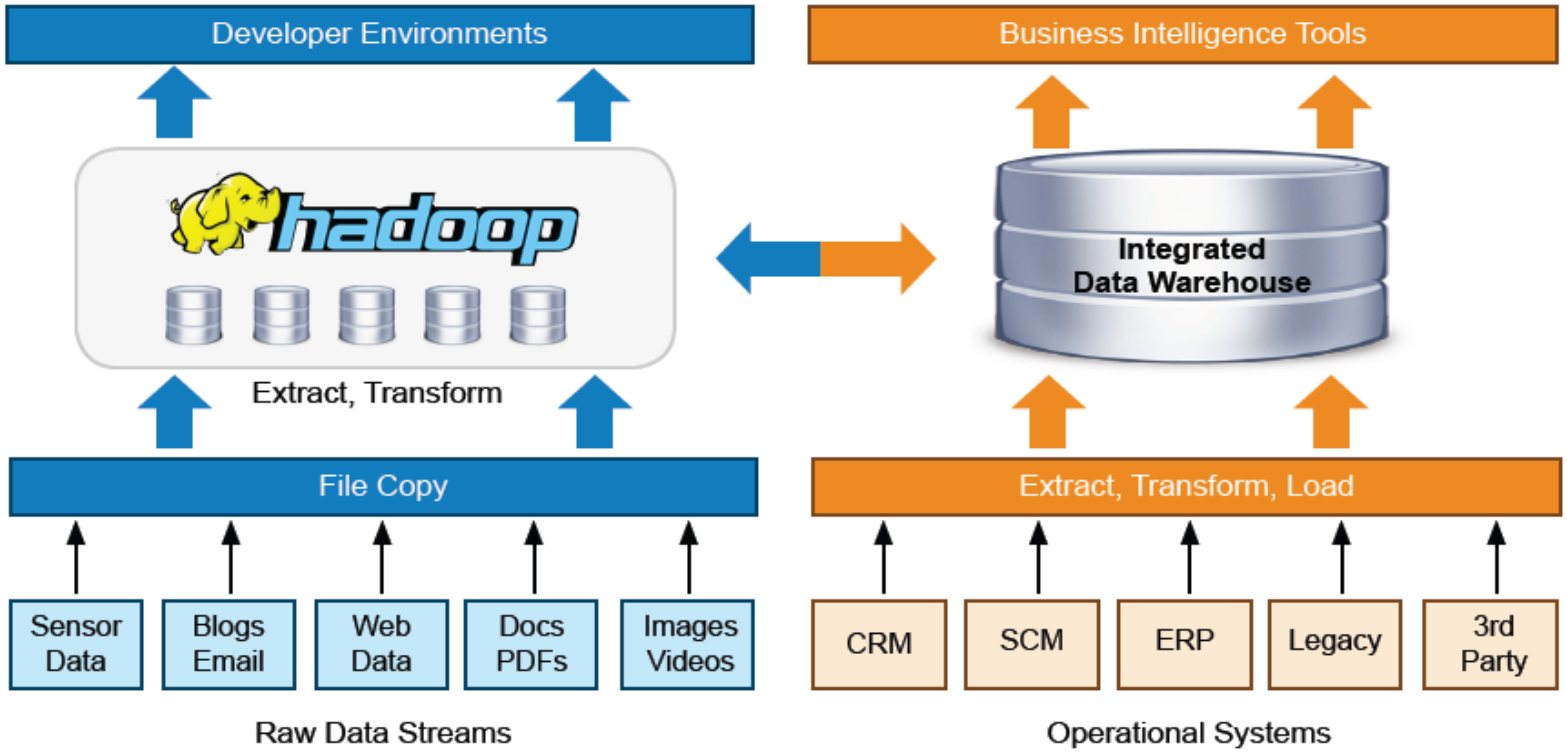

Telecom
Big Data
PM
Data Warehouse
Architecture
PM
Big Data
Architecture
Data Lake
Evolution

Big Data Platform Architecture
Online & Offline



Architecture Compare

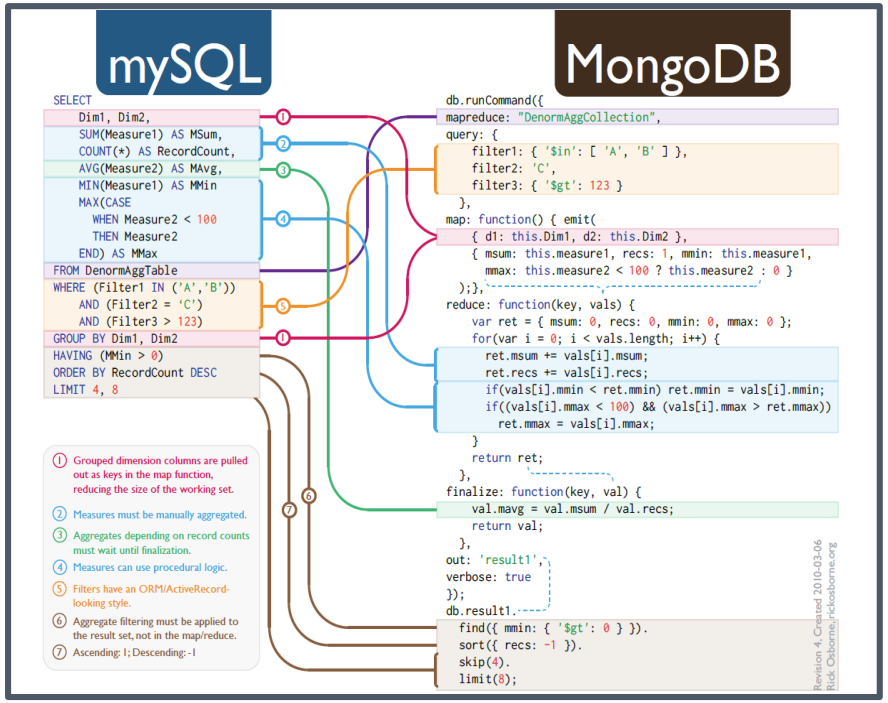
Same architecture for Data warehouse



NOKIA

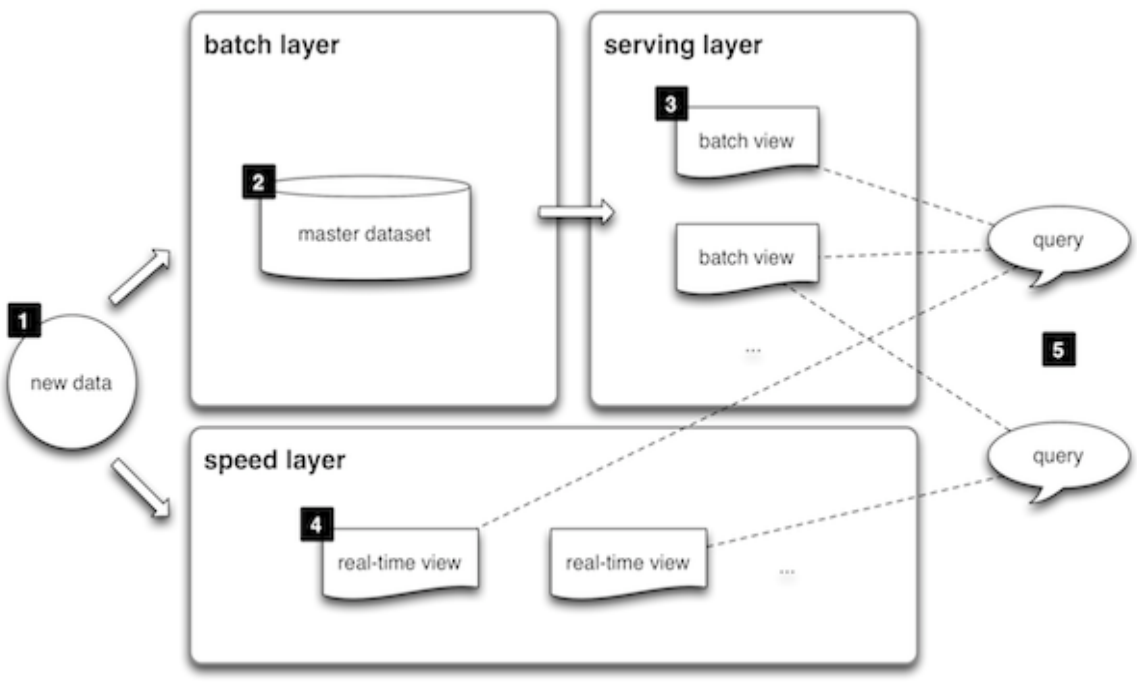
Architecture Compare

Group-By-Aggregation and Map-Reduce



- The Map-Reduce programming model provides a good abstraction of group-by-aggregation operations over a cluster of machines.
- The programmer provides a map function that performs grouping and a reduce function that performs aggregation.
- The underlying run-time system achieves parallelism by partitioning the data and processing different partitions concurrently using multiple machines.

Reference architecture
Lambda architecture is popular



speed layer
(i) compensates for the high latency of updates to the serving layer
(ii) deals with recent data only

serving layer
(i) indexes the batch views
(ii) Can be queried in low-latency, ad-hoc way

batch layer
(i) managing the master dataset (an immutable, append-only set of raw data),
(ii) pre-compute the batch views

NOKIA

Telecom
Big Data

PM
Data Warehouse
Architecture

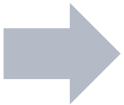
PM
Big Data
Architecture

Data Lake
Evolution

Flat operations architecture

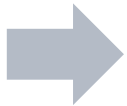
Data Lake Motivation

Some pain point for the hierarchy architecture :
Slow response times;
Poor scaling;
Slow development;
Overlapping development.

Flat operations architecture  Data Lake

NOKIA

Distributed analytics
Data Lake Motivation

Distributed analytics  Data Lake

NOKIA

5G Impact

Data Lake Motivation

5G → BigData (Semi-Structured and Unstructured) → Modern Data Architecture for Enterprise → Data Lake Storage Architecture → Data Lake

Data Lake

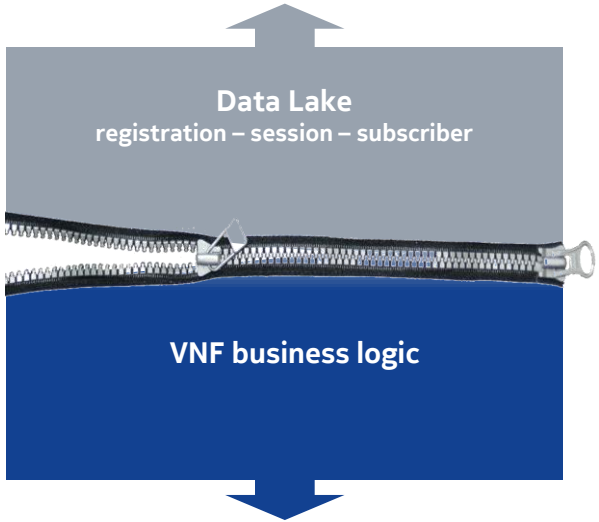
NOKIA

Cloud transformation

Data Lake Motivation

Cloud → Network Function Cloudification → Network Function Virtualization → stateless VNF→ Distributed Sharing Storage → Data Lake

Non-breakable, Open, Ultra-fast,
Service-logic agnostic,
Multi-tenant capable



Data Lake

Stateless, Scalable, Self-organizing

NOKIA

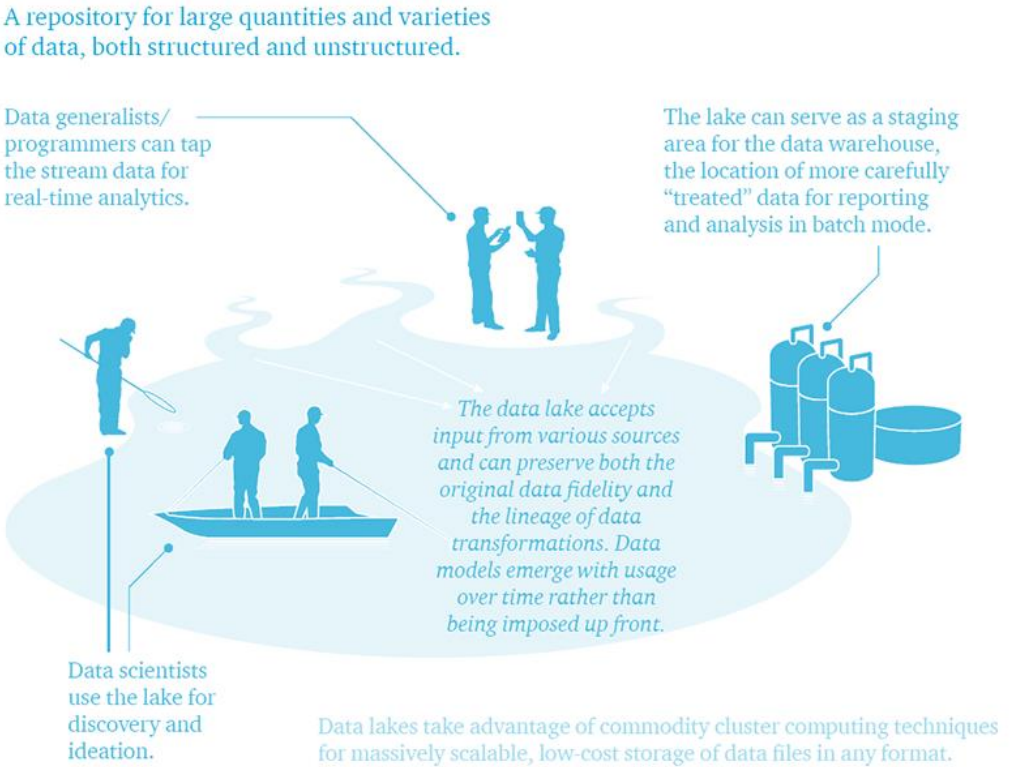
What is Data Lake?
PWC definition

"Size and low cost"

"Fidelity: Hadoop data lakes preserve data in its original form"

"Ease of accessibility: Accessibility is easy in the data lake"

"Late binding: Hadoop lends itself to flexible, task-oriented structuring and does not require up-front data models"



NOKIA

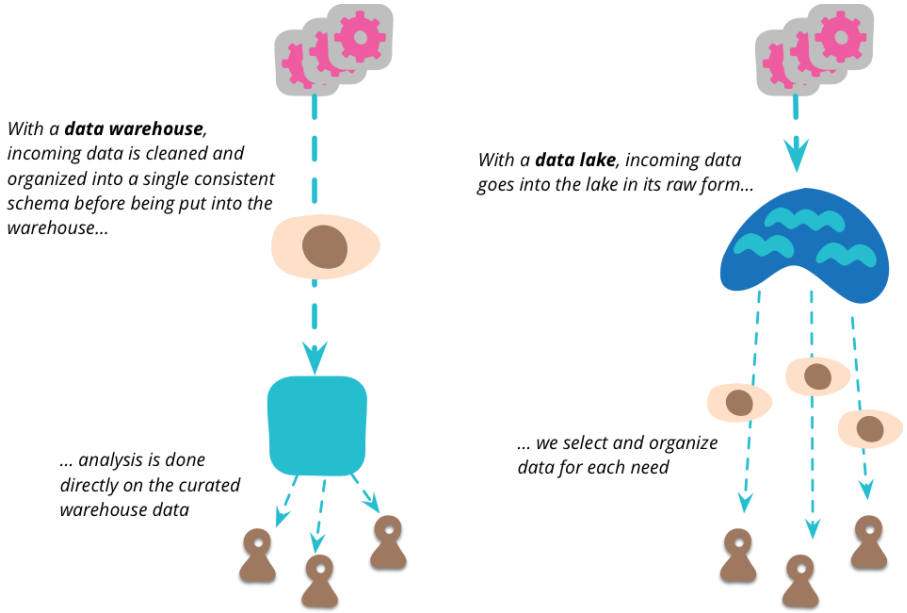
What is Data Lake?
Martin fowler ‘s view

Data warehouse:

- *It represents an abstracted picture of the business organized by subject area.*
- *It is highly transformed and structured.*
- *Data is not loaded to the data warehouse until the use for it has been defined.*
- *It generally follows a methodology such as those defined by Ralph Kimball and Bill Inmon.*

Data lake :

- *All data is loaded from source systems. No data is turned away.*
- *Data is stored at the leaf level in an untransformed or nearly untransformed state.*
- *Data is transformed and schema is applied to fulfill the needs of analysis.*



http://martinfowler.com/bliki/DataLake.html

NOKIA

What is Data Lake?

Comparing the Enterprise Data Warehouse and the Data Lake

Dimension	Enterprise Data Warehouse	Data Lake
Workload	Hundreds to thousands of concurrent users performing interactive analytics	Batch processing of data at scale.
Schema	Typically schema is defined before data is stored. **Schema on write** means required data is identified and modeled in advance.	Typically schema is defined after data is stored. **Schema on read** means data must be captured in code for each program accessing the data.
Data	**Cleansed**	**Raw**
Complexity	Complex joins	Complex processing
Scale	Can scale to large data volumes at moderate cost	Can scale to extreme data volumes at low cost
SQL	ANSI SQL, ACID compliant	Flexible programming, evolving SQL

NOKIA

Q&A

Thanks much for the participation!!!

