BIG DATA WITH NOT ONLY SQL

Philippe Julio

Open for Business...

WHO AM I

- Big Data / Analytics / BI & Cloud Solutions Specialist
- http://www.linkedin.com/in/JulioPhilippe
- Skills

BIG DATA MANAGEMENT INSIGHT

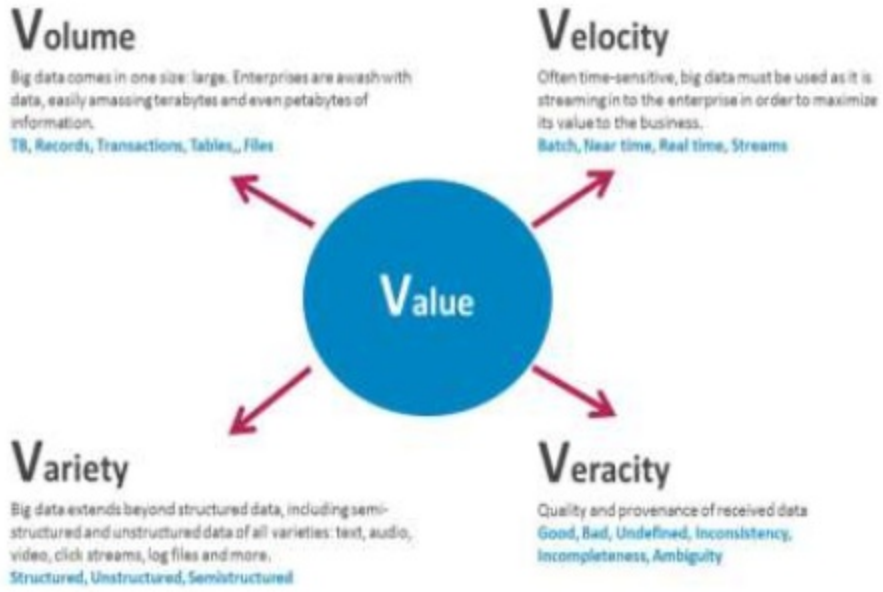
« Data don’t spring relevant, they become though ! »

DATA-DRIVEN ON-LINE WEBSITES

- To run the apps : messages, posts, blog entries, video clips, maps, web graph...
- To give the data context : friends networks, social networks, collaborative filtering...
- To keep the applications running : web logs, system logs, system metrics, database query logs...

BIG DATA – NOT ONLY DATA VOLUME

- Improve analytics and statistics models
- Extract business value by analyzing large volumes of multi-structured data from various sources such as databases, websites, blogs, social media, smart sensors...
- Have efficient architectures, massively parallel, highly scalable and available to handle very large data volumes up to several petabytes



Thematics

- Web Technologies
- Database Scale-out
- Relational Data Analytics
- Distributed Data Analytics
- Distributed File Systems
- Real Time Analytics

BIG DATA APPLICATIONS DOMAINS

- **Digital marketing optimization** (e.g., web analytics, attribution, golden path analysis)
- **Data exploration and discovery** (e.g., identifying new data-driven products, new markets)
- **Fraud detection and prevention** (e.g., revenue protection, site integrity & uptime)
- **Social network and relationship analysis** (e.g., influencer marketing, outsourcing, attrition prediction)
- **Machine-generated data analytics** (e.g., remote device insight, remote sensing, location-based intelligence)
- **Data retention** (e.g. long term conservation, data archiving

SOME BIG DATA USE CASES BY INDUSTRY

Energy

- Smart meter analytics
- Distribution load forecasting & scheduling
- Condition-based maintenance

Telecommunications

- Network performance
- New products & services creation
- Call Detail Records (CDRs) analysis
- Customer relationship management

Retail

- Dynamic price optimization
- Localized assortment
- Supply-chain management
- Customer relationship management

Manufacturing

- Supply chain management
- Customer Care Call Centers
- Preventive Maintenance and Repairs
- Customer relationship management

Banking

- Fraud detection
- Trade surveillance
- Compliance and regulatory
- Customer relationship management

Insurance

- Catastrophe modeling
- Claims fraud
- Reputation management
- Customer relationship management

Public

- Fraud detection
- Fighting criminality
- Threats detection
- Cyber security

Media

- Large-scale clickstream analytics
- Abuse and click-fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

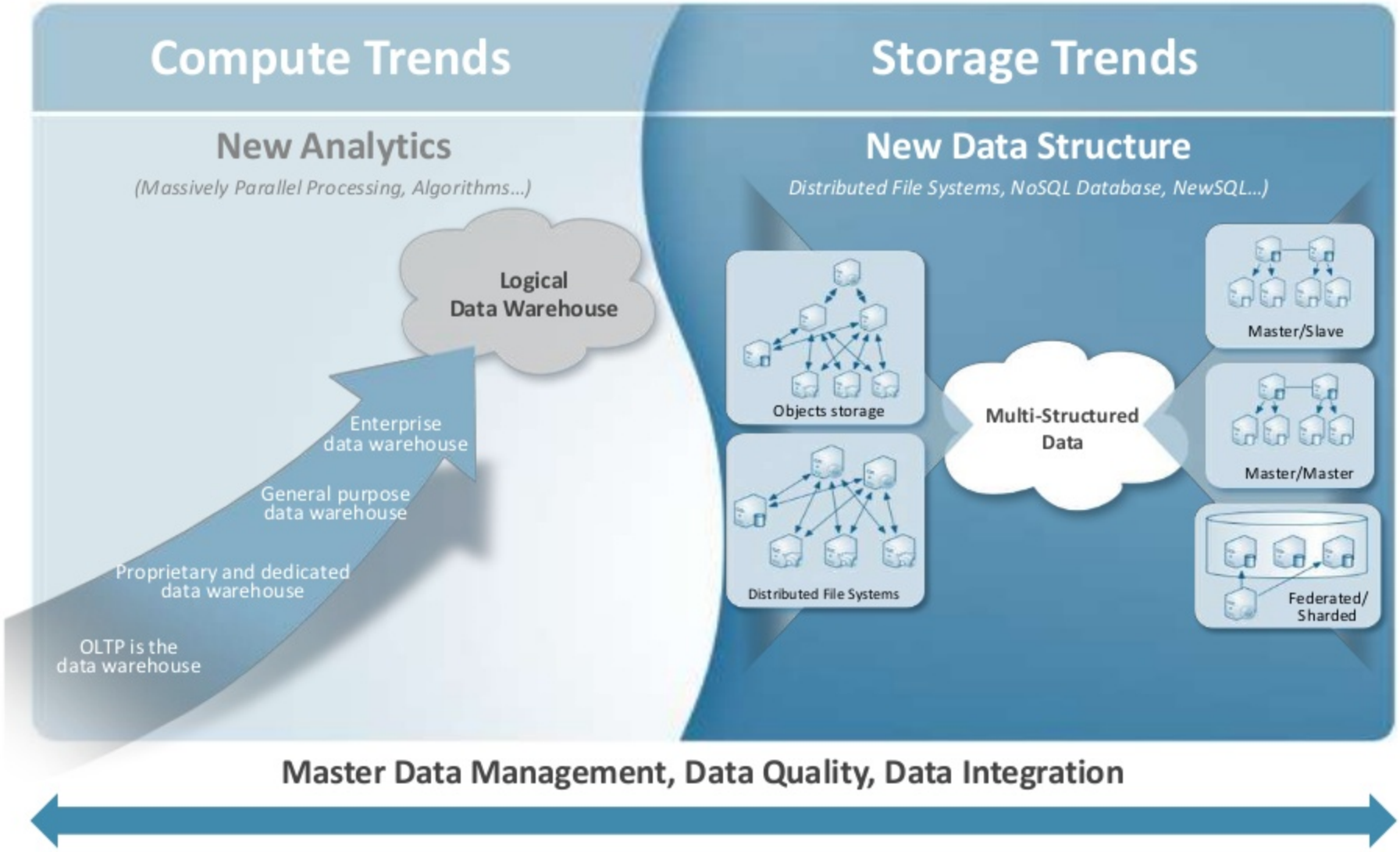
Healthcare

- Clinical trials data analysis
- Patient care quality and program analysis
- Supply chain management
- Drug discovery and development analysis

TOP 10 BIG DATA SOURCES

1. Social network profiles
2. Social influencers
3. Activity-generated data
4. SaaS & Cloud Apps
5. Public web information
6. MapReduce results
7. Data warehouse appliances
8. Columnar/NoSQL databases
9. Network and in-stream monitoring technologies
10. Legacy documents

NEW DATA AND MANAGEMENT ECONOMICS

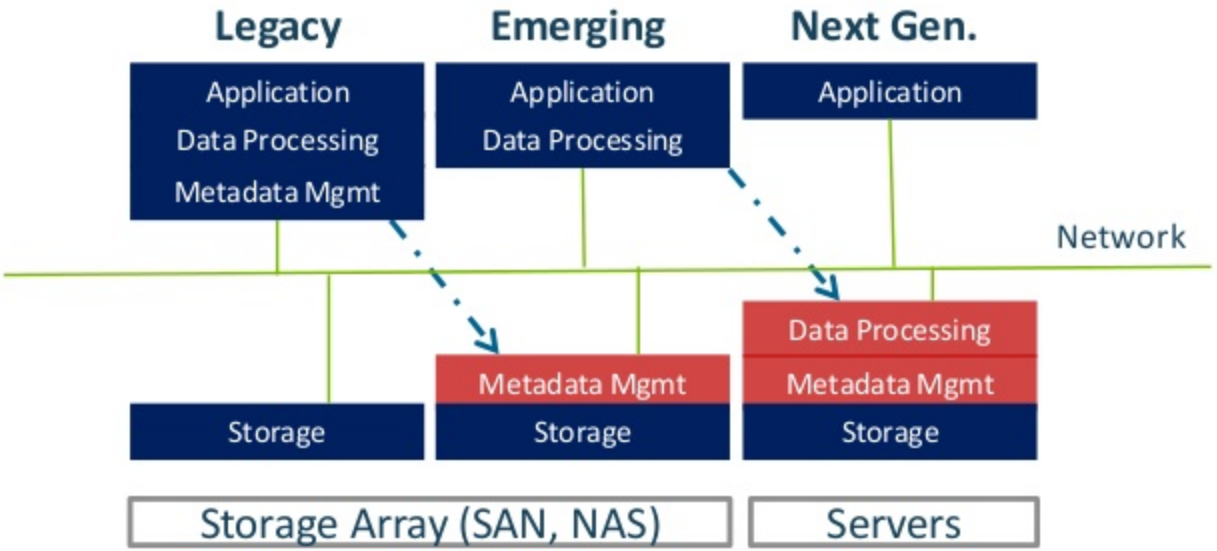


MOVING COMPUTATION TO STORAGE

General Purpose Storage Servers

- Combine server with disks & networking for reducing latency
- Specialized software enables general purpose systems designs to provide high performance data services

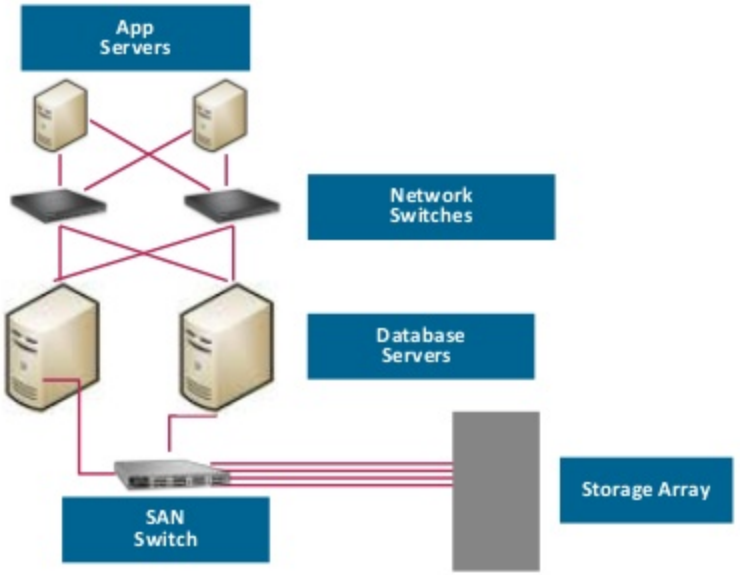
Moving Data processing to Storage



BIG DATA ARCHITECTURE

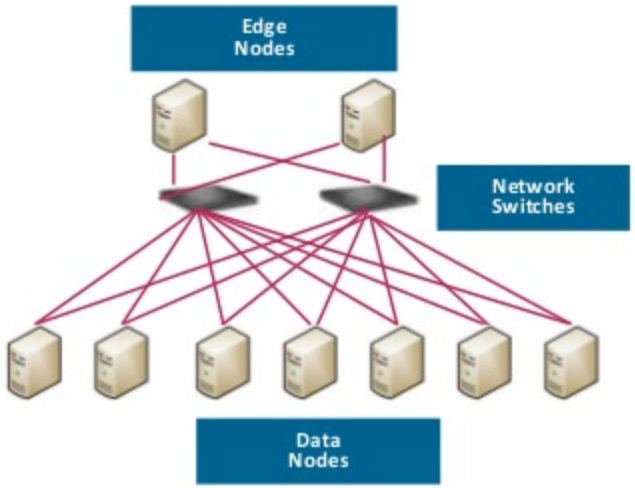
BI & DWH Architecture - Conventional

- SQL based
- High availability
- Enterprise database
- Right design for structured data
- Current storage hardware (SAN, NAS, DAS)



Analytics Architecture – Next Generation

- Not only SQL based
- High scalability, availability and flexibility
- Compute and storage in the same box for reducing the network latency
- Right design for semi-structured and unstructured data



DATA WAREHOUSE

- Data Warehouse appliances
  - EMC Greenplum
  - Microsoft Parallel Data Warehouse
  - IBM Netezza
  - Oracle Exadata
  - SAP HANA
  - ParAccel Analytic Database
  - Teradata
  - HP Vertica
- SQL Database
- Massively Parallel Processing
- Hadoop Connectivity
- Column-Oriented database
- In-Memory database

MAPREDUCE ALGORITHMS

MapReduce

- MapReduce is the programming paradigm popularized by Google researchers
- Open-source Hadoop implementation of MapReduce by Yahoo
- Open source software framework for distributed computation
- Parallel computation (Map) on each block (Split) of data in an DFS file and output a stream of (Key, Value) pairs to the local file system
- JobTracker schedules and manages jobs
- TaskTracker executes individual map() and reduce() tasks on each cluster node

Algorithms

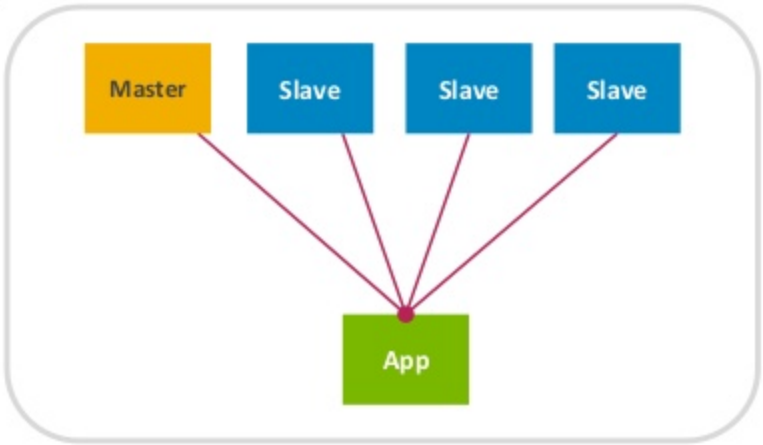
- Association Rule Learning Algorithms
- Genetic Algorithms
- Neural Network Algorithms
- Statistical Algorithms (Pandas)
- Machine Learning Algorithms (Mahout, Weka, Scikit Learn)
- Natural Language Processing Algorithms
- Trading Algorithms
- Clinical design Algorithms
- Searching Algorithms (Lucene, Solr, Katta, ElasicSearch, OpenSearchServer…)

Languages

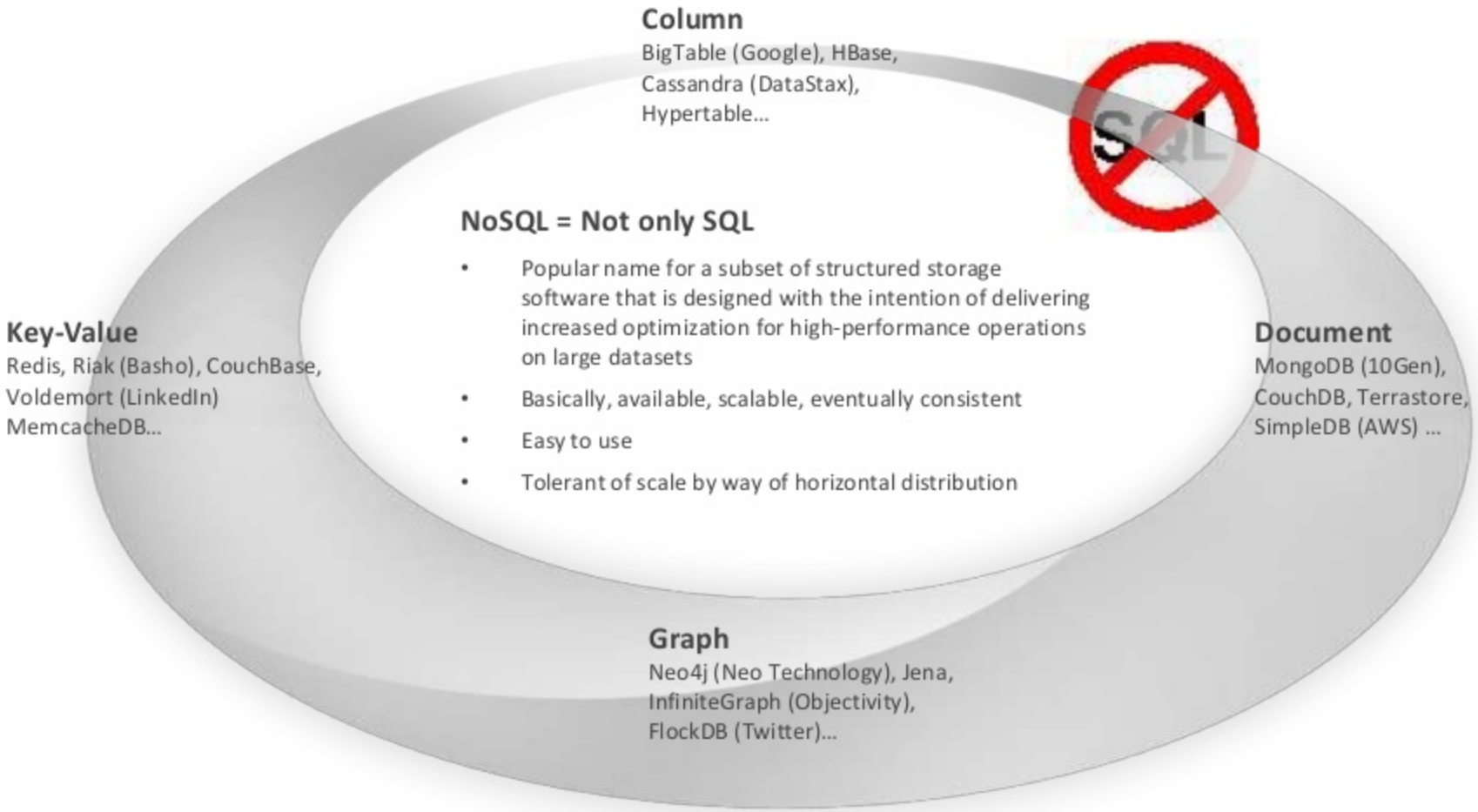
- PHP
- Erlang
- Python
- Ruby
- R
- Java

DISTRIBUTED FILE SYSTEMS

- System that permanently store data
- Divided into logical units
  (files, shards, chunks, blocks…)
- A file path joins file and directory names into a relative or absolute address to identify a file
- Support access to file and remote servers
- Support concurrency
- Support distribution
- Support replication
- NFS, GPFS, Hadoop
  HDFS, GlusterFS, MogileFS, MooseFS….



NOSQL DATABASES CATEGORIES



NOSQL DATABASES CATEGORIES

Key-Value

- Store items as alphanumeric identifier (Key)
- Associate values in a simple standalone tables
- Values must be (string, list, set)
- Data search base on key
- Fast and highly scalable to retrieve a value
- Domains: managing user profiles, retrieving product name…

Column

- BigTable-style database
- Column-oriented data structure that accommodates multiple attributes per key
- Petabyte scale
- Domains: Distributed data storage, Versioning with timestamp, Sorting, Parsing
- Data exploration

Document

- Documents (objects) map nicely to programming language data types
- Value = Collection>Document>Field
- Embedded documents and arrays reduce need for joins
- Dynamically-typed for easy schema evolution
- No joins and no multi-document transactions for high performance and easy scalability

Graph

- Structured relational graphs of interconnected key-value pairings
- Object-oriented network of nodes (Node), Nodes Relationship (Edge), Properties (nodes attributes expressed as key-value pairs)
- Relation between data
- Domains: social networks, recommendations, investigations, relationships…

Key	Value
User001	Peter
User002	Paul
User003	Rick

Key	Timestamp	Type	Size
E1	12	Zebra	Medium
	11	Lion	Big
E2	13	Bird	Small

Collection

Document	Name	Age
Doc001	Paul	30
Doc002	Jacques	35

Node

Node	Name	Age
X	John	30
Y	Bob	50

Edge

a	b
X	Y
Y	X

NoSQL Data Modeling Techniques

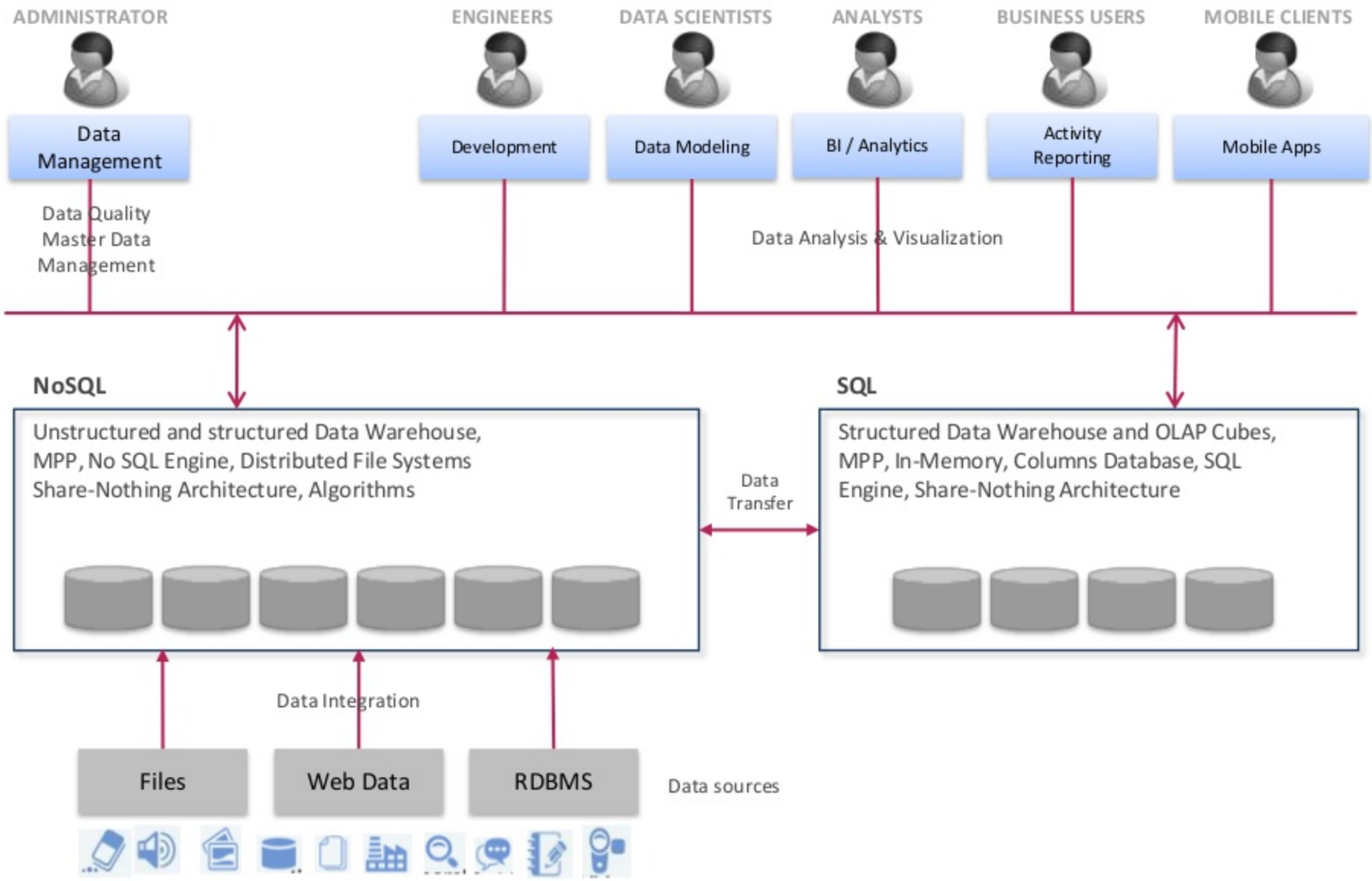
Geo hashing, Index table, Composite keys aggregation, Materialized paths…

http://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques/

NEW SQL

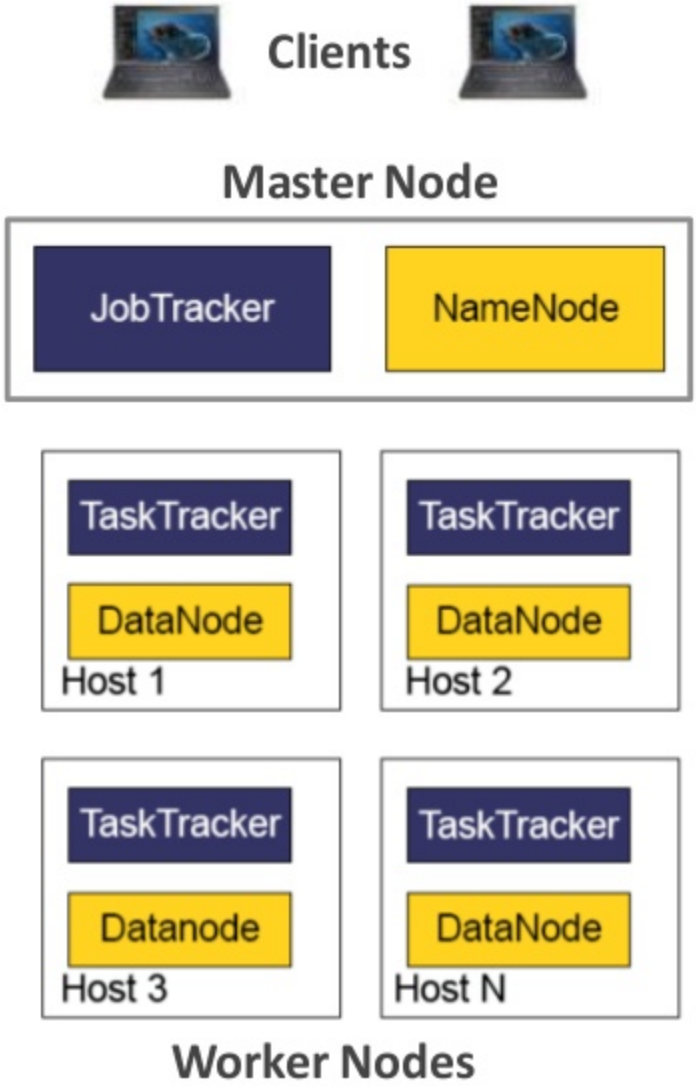
- Relational database with horizontal scalability
- MySQL Ecosystem
- Distributed database with MySQL compliance: Cubrid
- Analytic database: InfiniDB
- In-Memory database with MySQL compliance: VoltDB

BIG DATA ARCHITETURE OVERVIEW



HDFS & MAPREDUCE

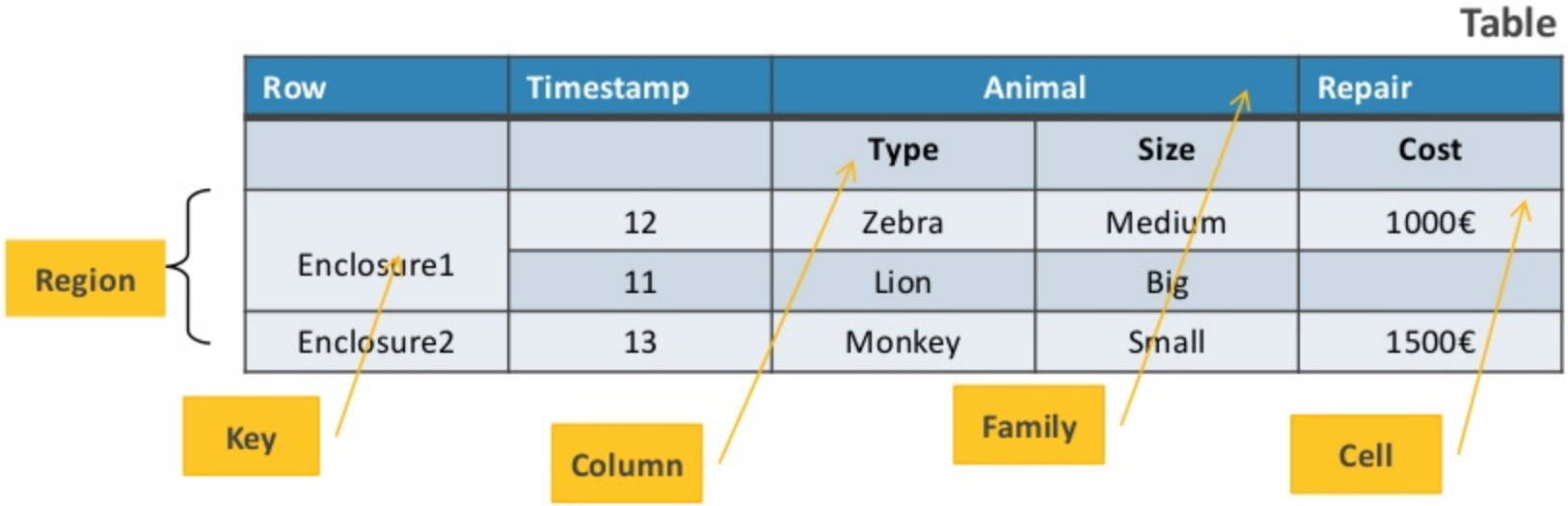
- **Hadoop Distributed File System**
  - A scalable, Fault tolerant, High performance distributed file system
  - Asynchronous replication
  - Write-once and read-many (WORM)
  - Hadoop cluster with 3 DataNodes minimum
  - Data divided into blocks, each block replicated 3 times (default)
  - No RAID required for DataNode
  - Interfaces: Java, Thrift, C Library, FUSE, WebDAV, HTTP, FTP
  - **NameNode** holds filesystem metadata
  - Files are broken up and spread over the **DataNodes**
- **Hadoop MapReduce**
  - Software framework for distributed computation
  - Input | Map() | Copy/Sort | Reduce() | Output
  - **JobTracker** schedules and manages jobs
  - **TaskTracker** executes individual map() and reduce() tasks on each cluster node



HBASE

- Clone of Big Table (Google)
- Implemented in Java (Clients : Java, C++, Ruby...)
- Data is stored "Column-oriented"
- Distributed over many servers
- Tolerant of machine failure
- Layered over HDFS
- Strong consistency

- It's not a relational database (No joins)
- Sparse data – nulls are stored for free
- Semi-structured or unstructured data
- Data changes through time
- Versioned data
- Scalable – Goal of billions of rows x millions of columns



(Table, Row_Key, Family, Column, Timestamp) = Cell (Value)