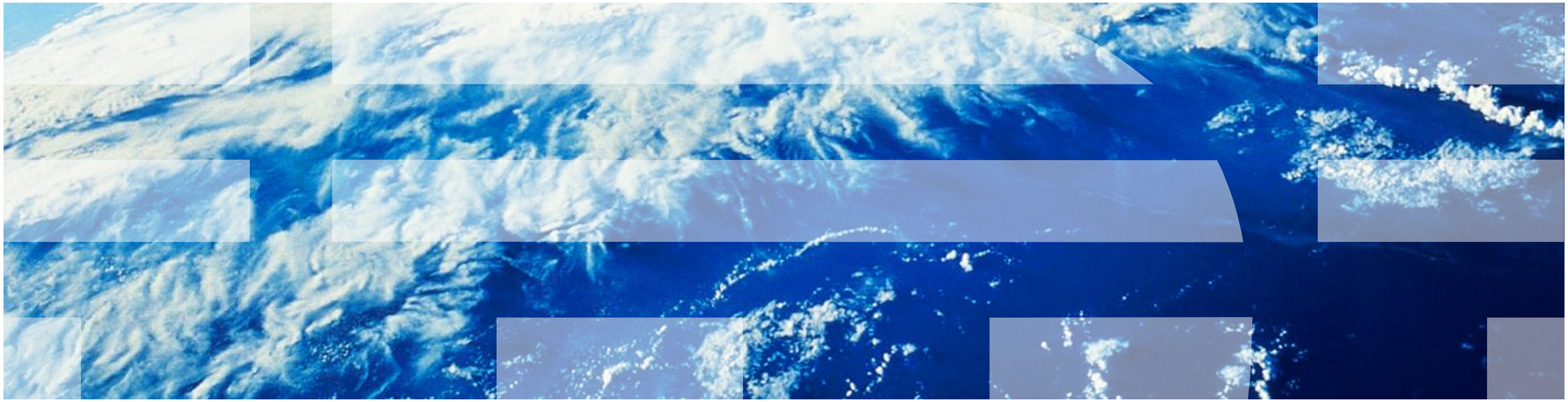
Central
Ohio
DB2
Users
Group

Modernizing the Data Warehouse

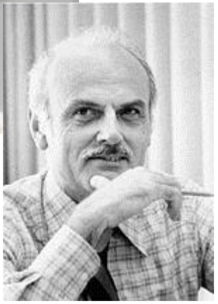
The Marriage of Big Data and Relational Technologies



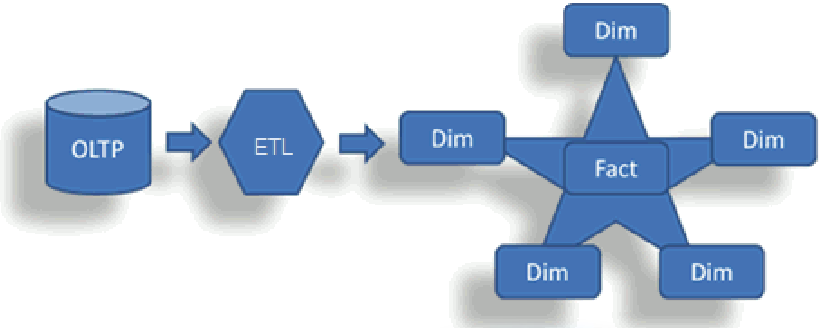
Dirk deRoos
dderoos@ca.ibm.com
@Dirk_deRoos
World-Wide Technical Sales, Big Data

The Evolution of Analytics

- **1960s: Navigational DBMS**
  – IMS (hierarchical)
- **1970s-1980s: Relational DBMS**
  – SQL
  – System R, System Z, DB2
- **1990s: Data Warehouse**
  – Dimensional model, ETL, MDM
- **Today: Big Data/NoSQL**

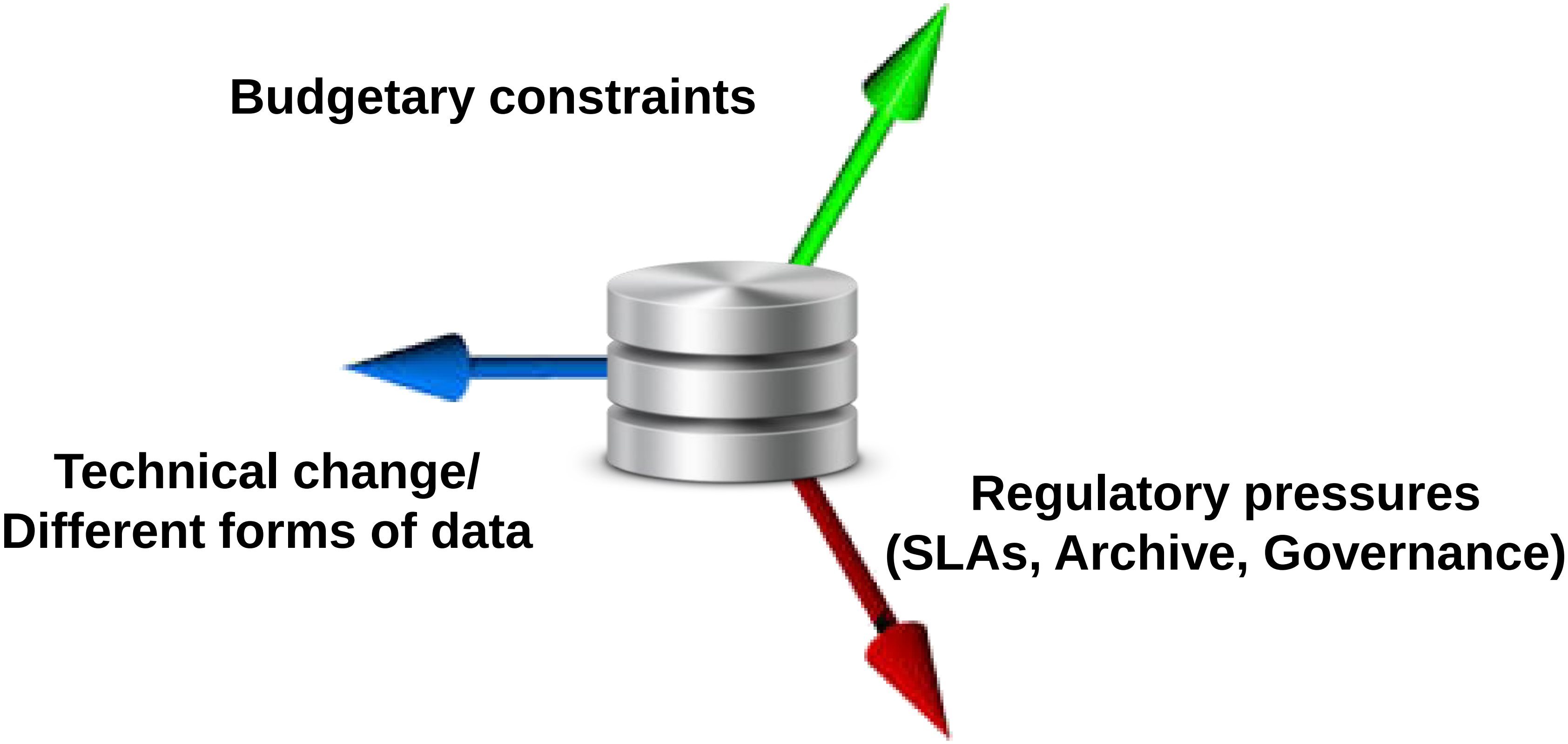


Ted Codd



© 2014 IBM Corporation

Pressures on Traditional Relational Stores



Budgetary constraints

Technical change/
Different forms of data

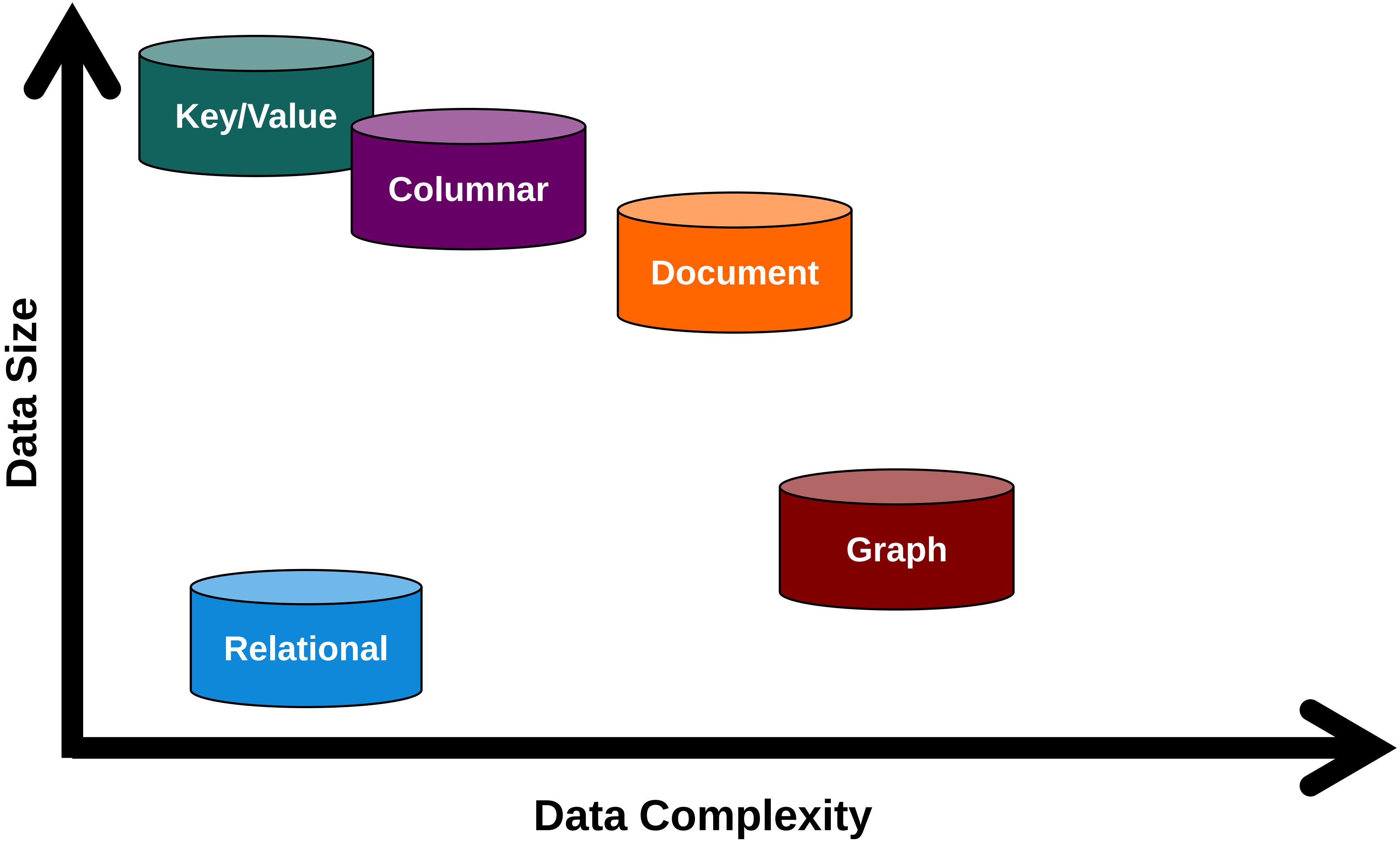
Regulatory pressures
(SLAs, Archive, Governance)

The NoSQL Revolution

- **Different requirements require different tools**
  - Document stores
  - Key/value stores
  - Google BigTable implementations
  - Graph databases
- **Values (there are exceptions)**
  - Huge data volumes – easy scale-out
  - Semi-structured data
  - Extreme performance

Database Genres

A High-level View



Traditional Warehousing vs. NoSQL

ACID vs. BASE

- Atomicity
- Consistency
- Isolation
- Durability

- Basically Available
- Soft state
- Eventually consistent

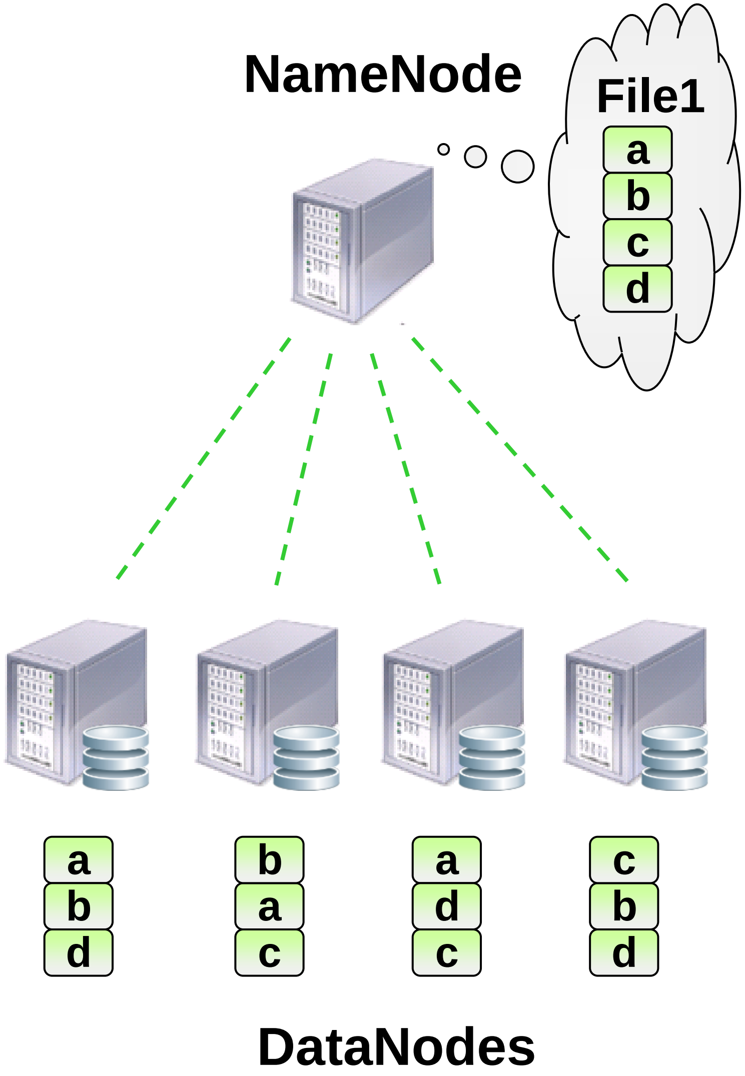
© 2014 IBM Corporation

Hadoop – Architecture

- **Master / Slave architecture**

- **Master: NameNode**
  - Manages the file system namespace and metadata
    - FsImage
    - EditLog
  - Regulates access by files by clients

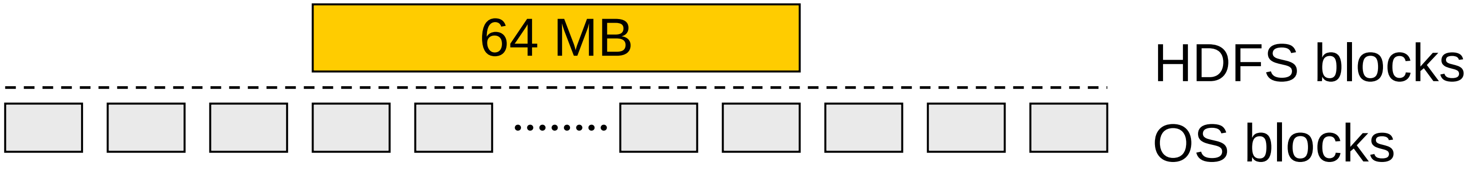
- **Slave: DataNode**
  - Many DataNodes per cluster
  - Manages storage attached to the nodes
  - Periodically reports status to NameNode
  - Data is stored across multiple nodes
  - Nodes and components will fail, so for reliability data is replicated across multiple nodes



Hadoop Distributed File System

- **HDFS is designed to support very large files**
- **Each file is split into blocks**
  - Hadoop default: 64MB
  - BigInsights default: 128MB

- **Blocks reside on different physical DataNode**
- **Behind the scenes, 1 HDFS block is supported by multiple operating system blocks**



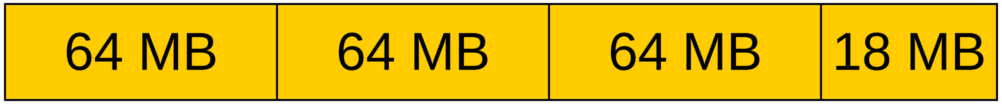
64 MB

HDFS blocks

....

OS blocks

- **If a file or a chunk of the file is smaller than the block size, only needed space is used. E.g.: a 210MB file is split as follows:**

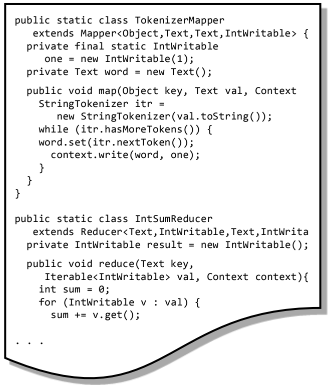


64 MB | 64 MB | 64 MB | 18 MB

MapReduce Explained

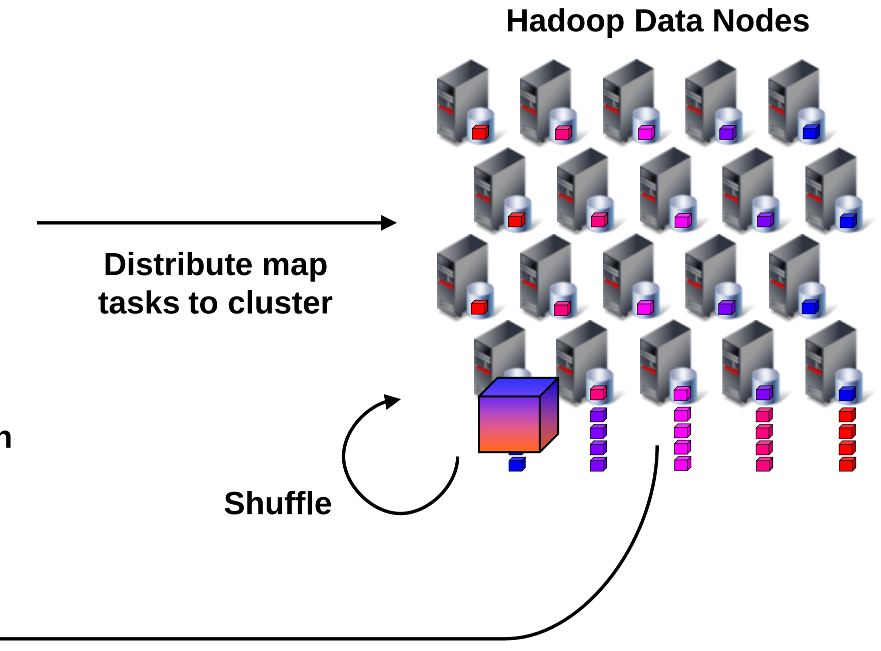
- **Hadoop computation model**
  - Data stored in a distributed file system spanning many inexpensive computers
  - Bring function to the data
  - Distribute application to the compute resources where the data is stored

- **Scalable to thousands of nodes and petabytes of data**



```
public static class TokenizerMapper
   extends Mapper<Object,Text,Text,IntWritable> {
  private final static IntWritable
     one = new IntWritable(1);
  private Text word = new Text();

  public void map(Object key, Text val, Context
    StringTokenizer itr =
       new StringTokenizer(val.toString());
    while (itr.hasMoreTokens()) {
    word.set(itr.nextToken());
      context.write(word, one);
    }
  }
}

public static class IntSumReducer
   extends Reducer<Text,IntWritable,Text,IntWrita
  private IntWritable result = new IntWritable();

  public void reduce(Text key,
     Iterable<IntWritable> val, Context context){
    int sum = 0;
    for (IntWritable v : val) {
      sum += v.get();

. . .
```

MapReduce Application



Hadoop Data Nodes

Distribute map tasks to cluster

Shuffle

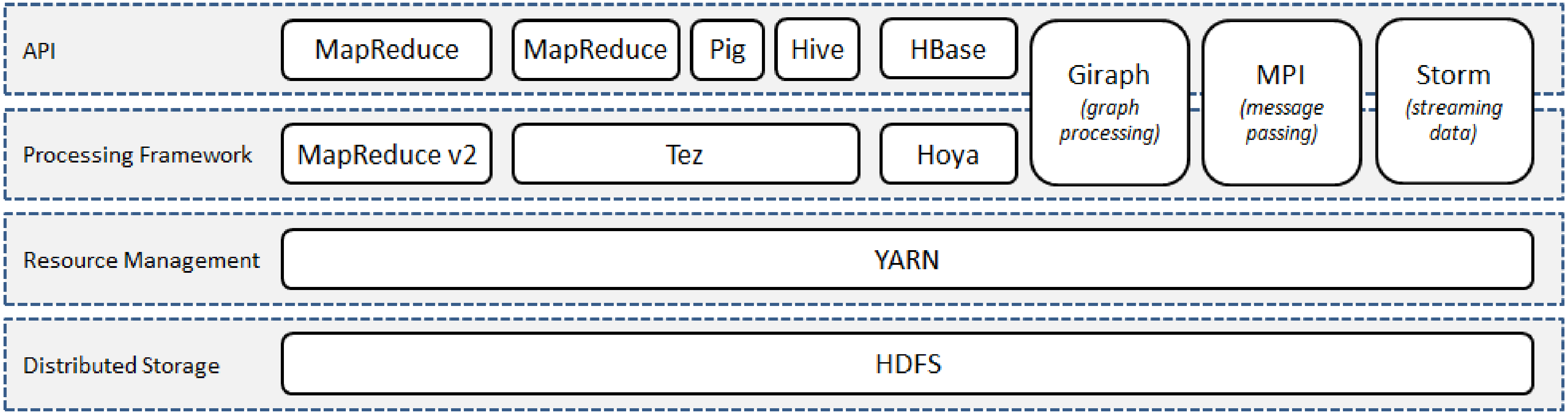
Result Set

Return a single result set

1. Map Phase
   (break job into small parts)
2. Shuffle
   (transfer interim output for final processing)
3. Reduce Phase
   (boil all output down to a single result set)

Next Generation Hadoop

- **Beyond MapReduce**
- **General purpose storage and processing framework**



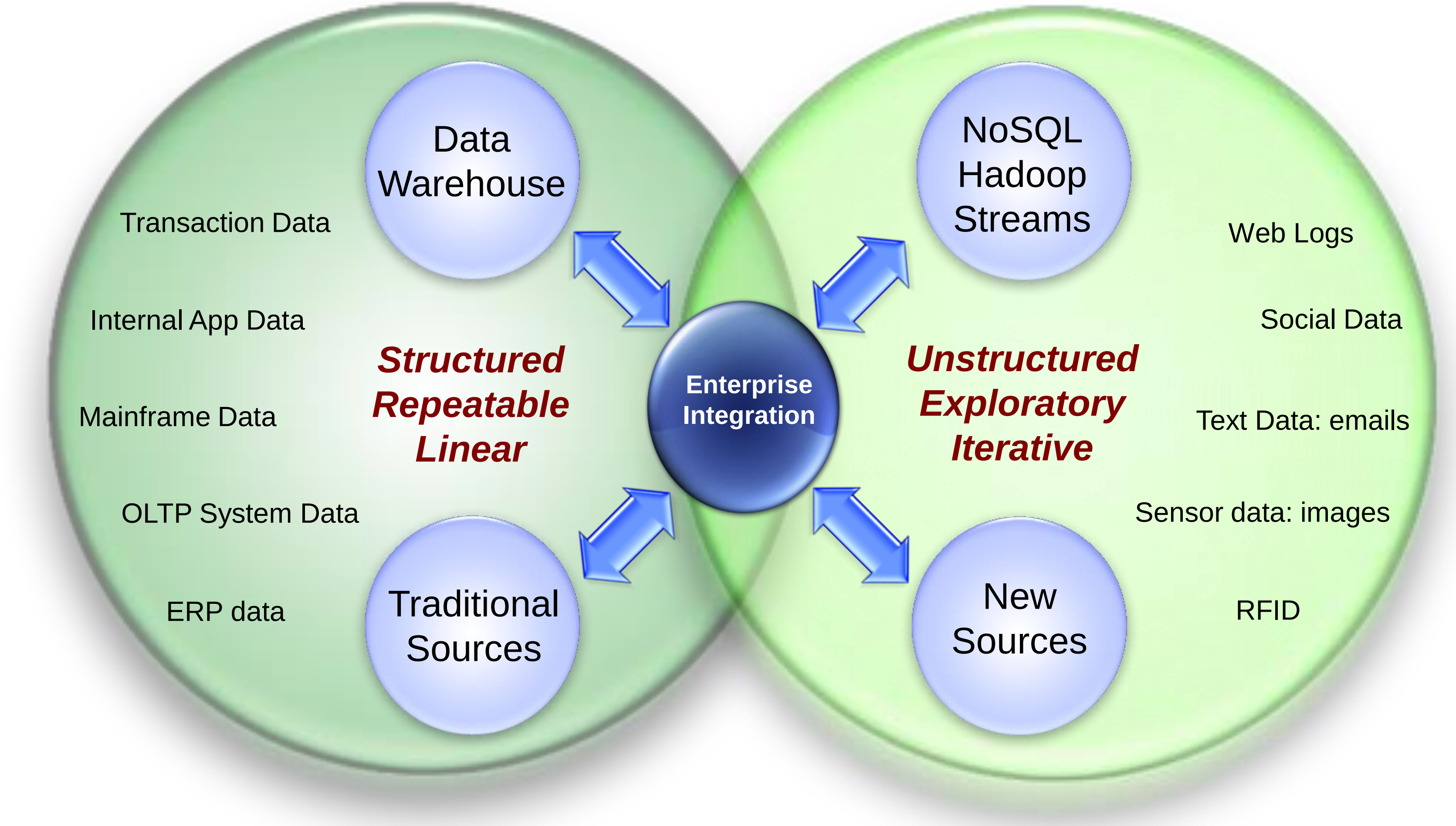
Complementary Analytics

Traditional Approach

Structured, analytical, logical

New Approach

Creative, holistic thought, intuition



Data Warehouse

Traditional Sources

Transaction Data

Internal App Data

Mainframe Data

OLTP System Data

ERP data

Structured
Repeatable
Linear

Enterprise Integration

NoSQL
Hadoop
Streams

New Sources

Unstructured
Exploratory
Iterative

Web Logs

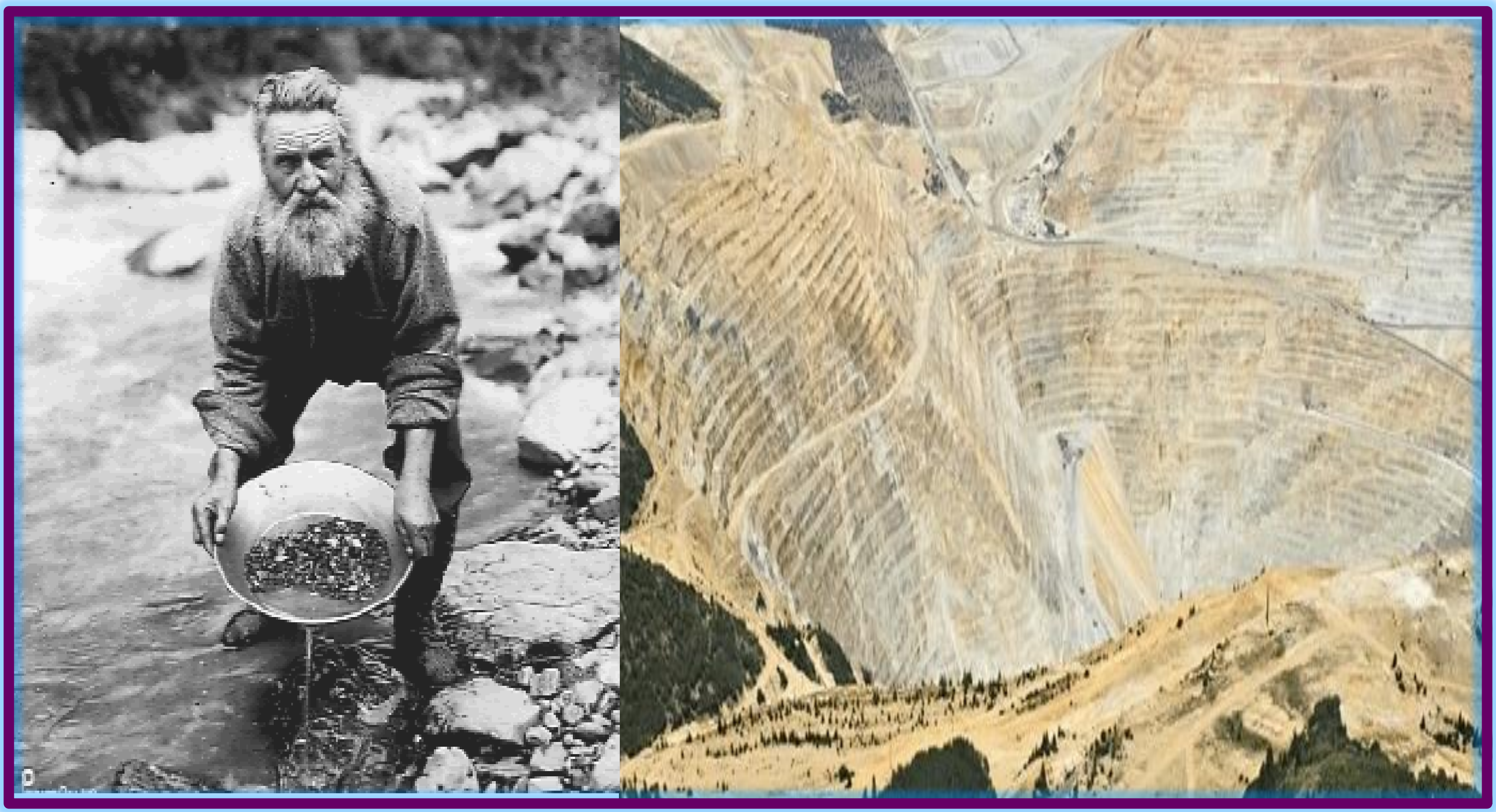
Social Data

Text Data: emails

Sensor data: images

RFID

Traditional Data Mining and Exploratory Analysis



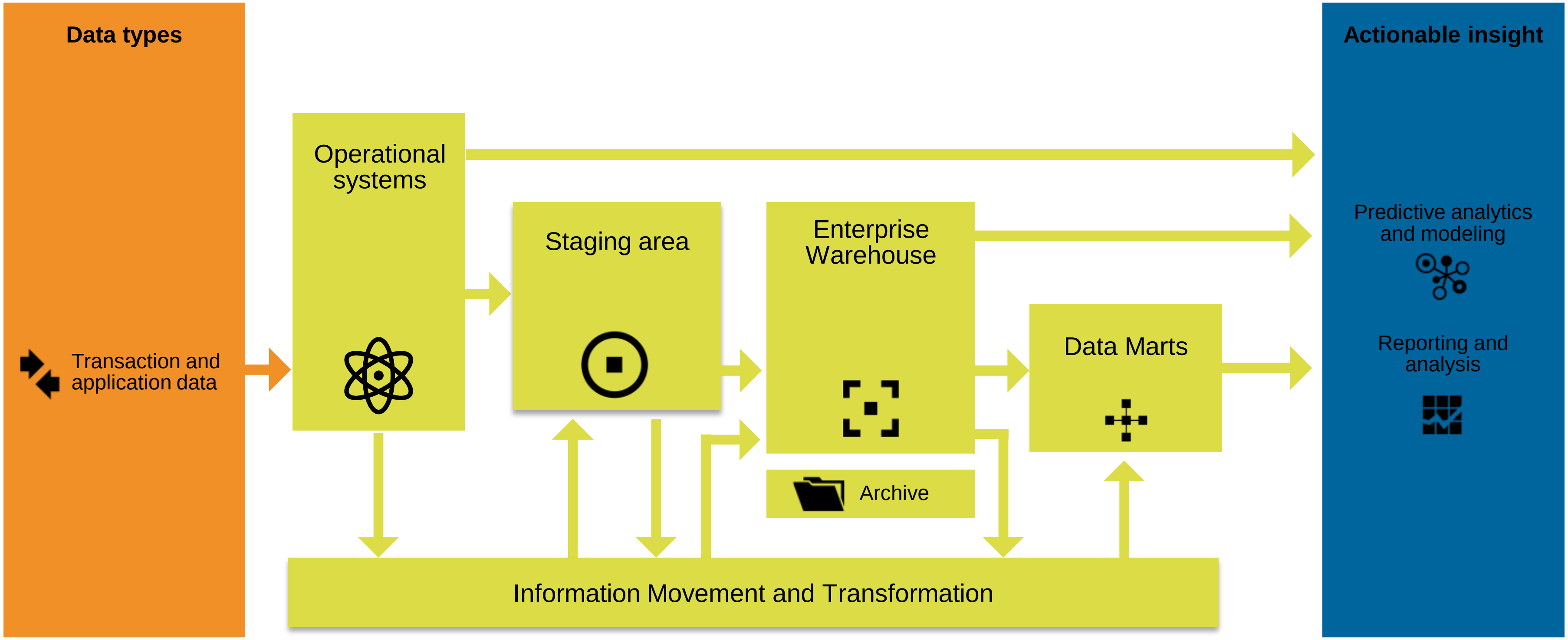
Data Governance Maturity Disciplines

- Organizational awareness
- Stewardship
- Policy
- Value creation
- Data risk management
- Security/Privacy/Compliance
- Data architecture
- Data quality
- Business glossary/metadata
- Information lifecycle management
- Audit and reporting

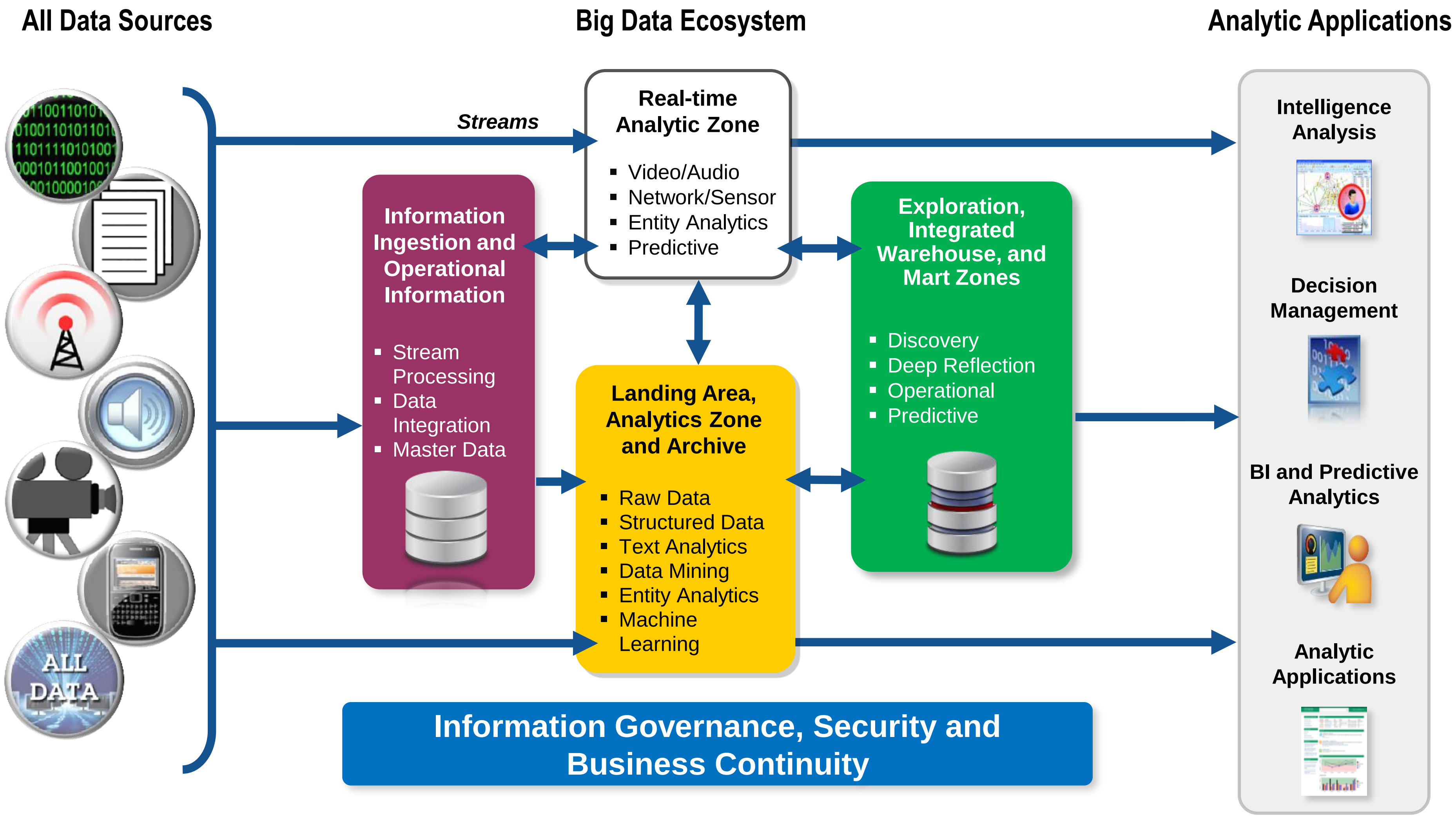
Data Governance Maturity Disciplines NoSQL Challenges

- Organizational awareness
- Stewardship
- Policy
- Value creation
- Data risk management
- Security/Privacy/Compliance
- Data architecture
- Data quality
- Business glossary/metadata
- Information lifecycle management
- Audit and reporting

Traditional Analytics



IBM Big Data Architecture Vision



All Data Sources

Big Data Ecosystem

Analytic Applications

Streams

Real-time Analytic Zone

- Video/Audio
- Network/Sensor
- Entity Analytics
- Predictive

Information Ingestion and Operational Information

- Stream Processing
- Data Integration
- Master Data

Landing Area, Analytics Zone and Archive

- Raw Data
- Structured Data
- Text Analytics
- Data Mining
- Entity Analytics
- Machine Learning

Exploration, Integrated Warehouse, and Mart Zones

- Discovery
- Deep Reflection
- Operational
- Predictive

Information Governance, Security and Business Continuity

Intelligence Analysis

Decision Management

BI and Predictive Analytics

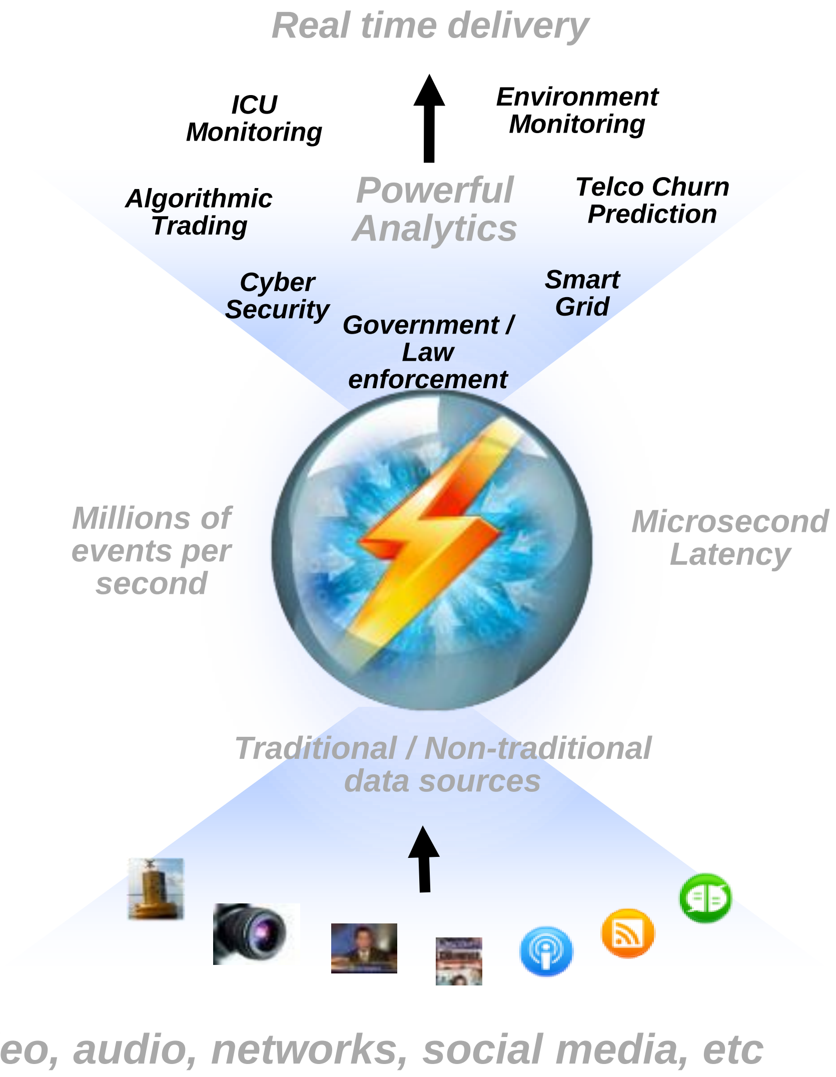
Analytic Applications

Analytics for Data-in-Motion

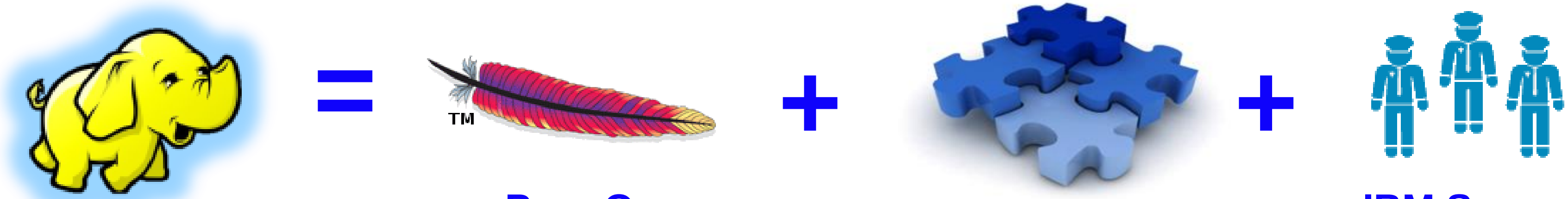
- **Scale-out architecture for massive linear scalability**

- **Sophisticated analytics with pre-built toolkits & accelerators**

- **Comprehensive development tools to build applications with minimal learning**



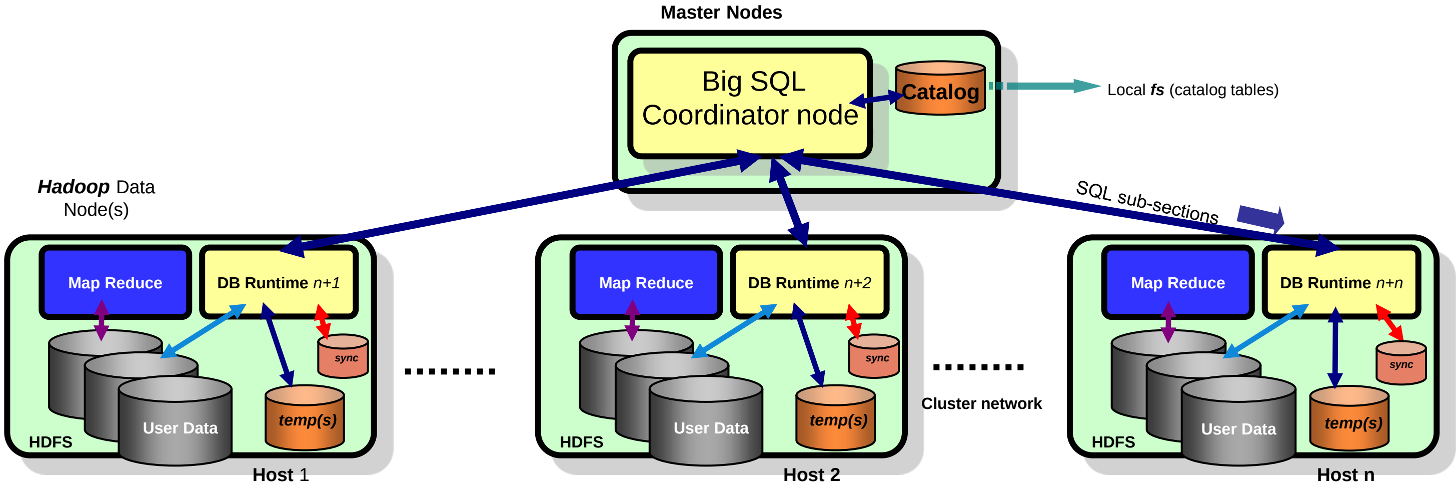
BigInsights: IBM’s Hadoop Distribution



BigInsights = Pure Open Source Code + Opt-in Enterprise Class Extensions + IBM Support Infrastructure

- **Analysis**
  - Native SQL interface
  - Native R interface
  - Text analysis toolkit
  - Social analysis toolkit
  - Spreadsheet style analysis GUI
- **Development lifecycle**
  - Cluster aware Eclipse plug-ins
  - App Store for Hadoop
- **Data Exploration**
  - Indexing and faceted search
  - Search-based applications
- **Management**
  - Enterprise file system
    - Advanced replication
    - Multi-temp storage
    - POSIX controls
  - Grid management
    - Mature resource manager
    - Multi-tenant workload support
- **Baked-in security**
  - LDAP
  - Role-based authorization
  - Perimeter security with reverse-proxy

Big SQL



- **Architecture**
  - IBM Optimizer + IBM Compiler + IBM Runtime => Ported to Hadoop
  - Nodes integrated in Hadoop cluster, direct access to Hadoop data
  - Queries Hadoop data – no proprietary data format
  - MapReduce run-time also available for query execution

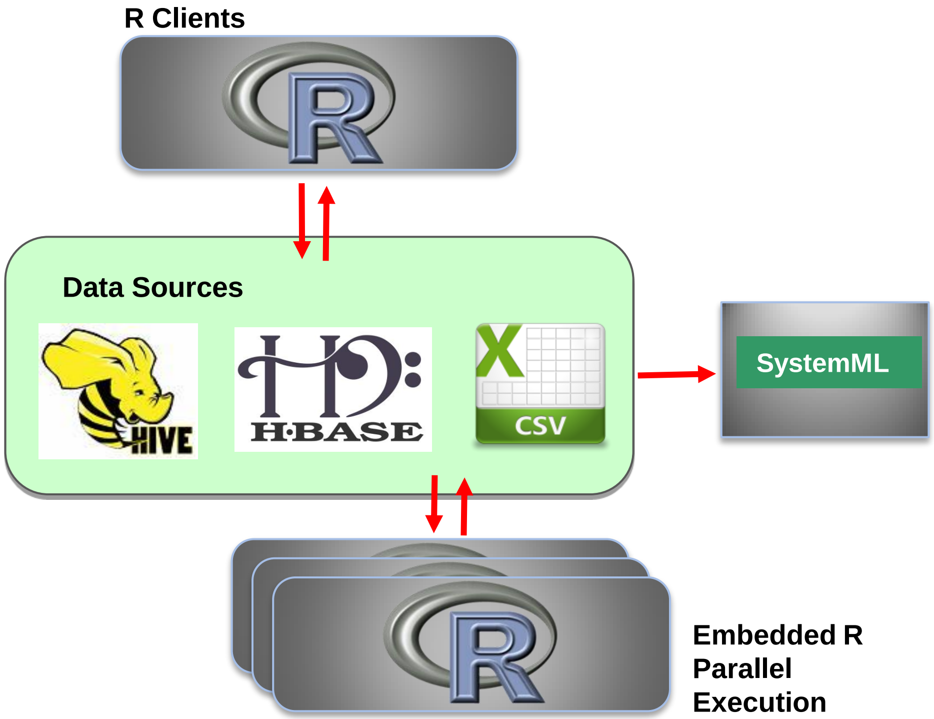
- **Benefits**
  - Extensive SQL support (ANSI, IBM, Oracle, Teradata)
  - Performance: Maturity – 30 years of engineering
  - Federated joins between relational systems and Hadoop
  - Security: Row and column access control

Deep Statistical Analysis: Big R

- **Fit-for-purpose architecture for deep statistical analysis**
  - Problems involving small data sets (10GB): R
  - Problems involving partitioned data sets (e.g. 32 x 10GB): BigR
  - Problems involving large data sets: (TB range): BigR using SystemML

- **R integration in BigInsights**
  - R code can be deployed against data stored in BigInsights
  - Big R: partitioning larger data sets and executing R code against them
  - Seamless access to data in BigInsights
  - Enterprise friendly license (no GPL)

- **SystemML**
  - Some data sets cannot be logically partitioned: too big for R
  - Engine designed for massive scale on Hadoop
  - Numerically accurate results
  - Provide an R interface for SystemML



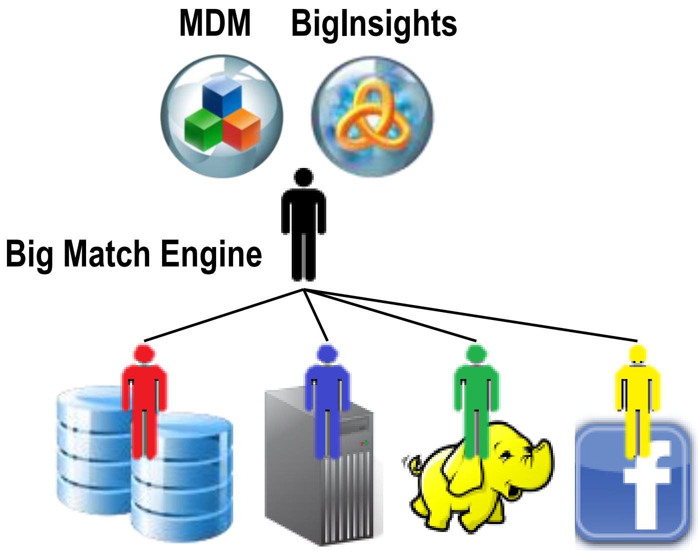
Big Match

Find and Integrate Master Data in Big Data Sources

- **How It Works**
  - Probabilistic matching on big data platform (BigInsights-Hadoop)
  - Matching at a higher volume
  - Matching of a wider variety of data sets

- **Client Value**
  - Find master data within big data sources
  - Get an answer faster – enable real-time matching at big data volumes

- **Building Big Data Confidence**
  - Provides more context by detecting master entities faster



Unique Data Matching Capabilities for Hadoop

Probabilistic matching engine and pre-built algorithms integrated into BigInsights for linking all data related to a customer natively within Hadoop

Internal / Structured

Web

Chris.johnson@cj.net

CRM

C. Johnson
123 Main Street
512-545-1234

Order Mgmt.

Chris Johnston
123 Main Street
512-554-1234
Shipping:
456 Pine Ave

Fulfillment

C. Johnson
125 Main Street
512-554-1234

Supply Chain

C. Johnson
Main Street
512-554-1234

Support Ticketing

Christine. Johnson
123 Main Street
Call length
Semi-structured notes
Satisfaction

Increased Value of Customer only if…

Big Match matches all these records

External / Unstructured

ChrisJohnson65
"Likes" Clothes,
Camping Gear

External Sources

Christine Johnson
Married
1 child
4/15/74

@ChristyJohnson65

3rd Party

Christy65
Mail Order responder
Specialty Apparel
Partner Sales data

Christy65
Circle / Network data

Predictive analytics and modeling

VIP: Gold
Customer Sat: 80%
Influence Score: 8/10

Match and Search Differentiators – Fuzzy Matching

- **Comprehensive library of fuzzy matching techniques**
- **Scored against probabilistic weights based on value frequencies in *your* data**

Phonetics

Mohammed vs. Mahmoud

Synonyms

Andrew = Andy
George = Jorge
1st = First

Abbreviations

AIG = American International Group
Road = Rd

Concatenation

Van de Velde = Vandevelde

Edit Distance

867-5309 ~ 876-5309

Transliteration

Toyota = トヨダ

Date Similarity

01/01/1973 ~ 01/02/1973

Proximity

Geocodes and great-circle distance

Typographical Errors

John Smith vs. John Snith

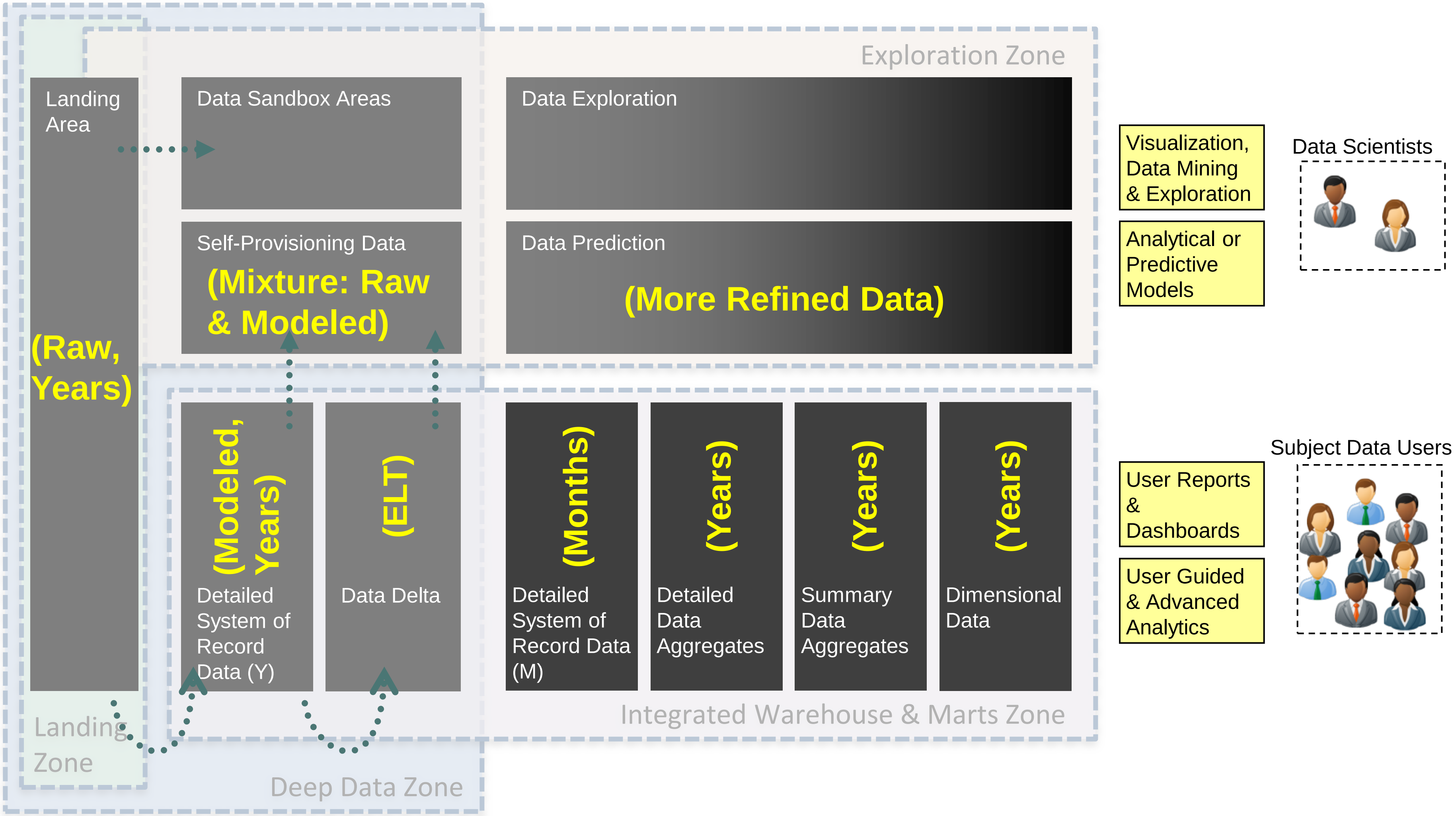
Noise Words

Initiate Inc. = Initiate

Misalignment

Kim Jung-il = Kim il Jung

Logical Data Warehouse – Schema Areas



Exploration Zone

Landing Area

(Raw, Years)

Data Sandbox Areas

Self-Provisioning Data

(Mixture: Raw & Modeled)

Data Exploration

Data Prediction

(More Refined Data)

(Modeled, Years)

Detailed System of Record Data (Y)

(ELT)

Data Delta

(Months)

Detailed System of Record Data (M)

(Years)

Detailed Data Aggregates

(Years)

Summary Data Aggregates

(Years)

Dimensional Data

Integrated Warehouse & Marts Zone

Landing Zone

Deep Data Zone

Visualization, Data Mining & Exploration

Analytical or Predictive Models

Data Scientists

User Reports & Dashboards

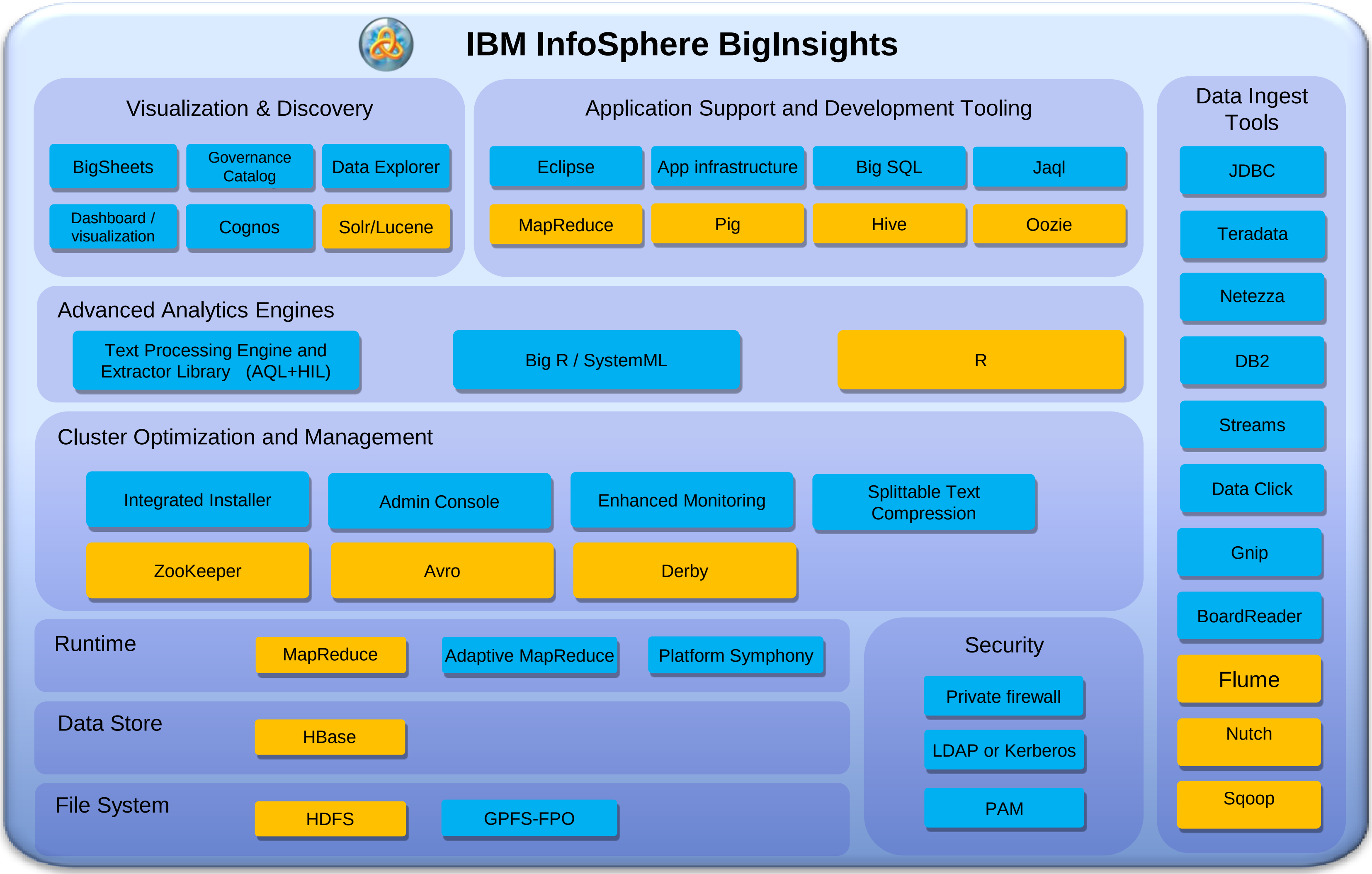
User Guided & Advanced Analytics

Subject Data Users

THINK

BIG

BigInsights Enterprise Edition Components



IBM InfoSphere BigInsights

Visualization & Discovery

- BigSheets
- Governance Catalog
- Data Explorer
- Dashboard / visualization
- Cognos
- Solr/Lucene

Application Support and Development Tooling

- Eclipse
- App infrastructure
- Big SQL
- Jaql
- MapReduce
- Pig
- Hive
- Oozie

Data Ingest Tools

- JDBC
- Teradata
- Netezza
- DB2
- Streams
- Data Click
- Gnip
- BoardReader
- Flume
- Nutch
- Sqoop

Advanced Analytics Engines

- Text Processing Engine and Extractor Library (AQL+HIL)
- Big R / SystemML
- R

Cluster Optimization and Management

- Integrated Installer
- Admin Console
- Enhanced Monitoring
- Splittable Text Compression
- ZooKeeper
- Avro
- Derby

Runtime

- MapReduce
- Adaptive MapReduce
- Platform Symphony

Data Store

- HBase

File System

- HDFS
- GPFS-FPO

Security

- Private firewall
- LDAP or Kerberos
- PAM

🟨	Open Source
🟦	IBM