HANA & Hadoop for Big Data Management

Will Gardella, Senior Director
SAP Applied Research - Big Data Program
william.gardella@sap.com

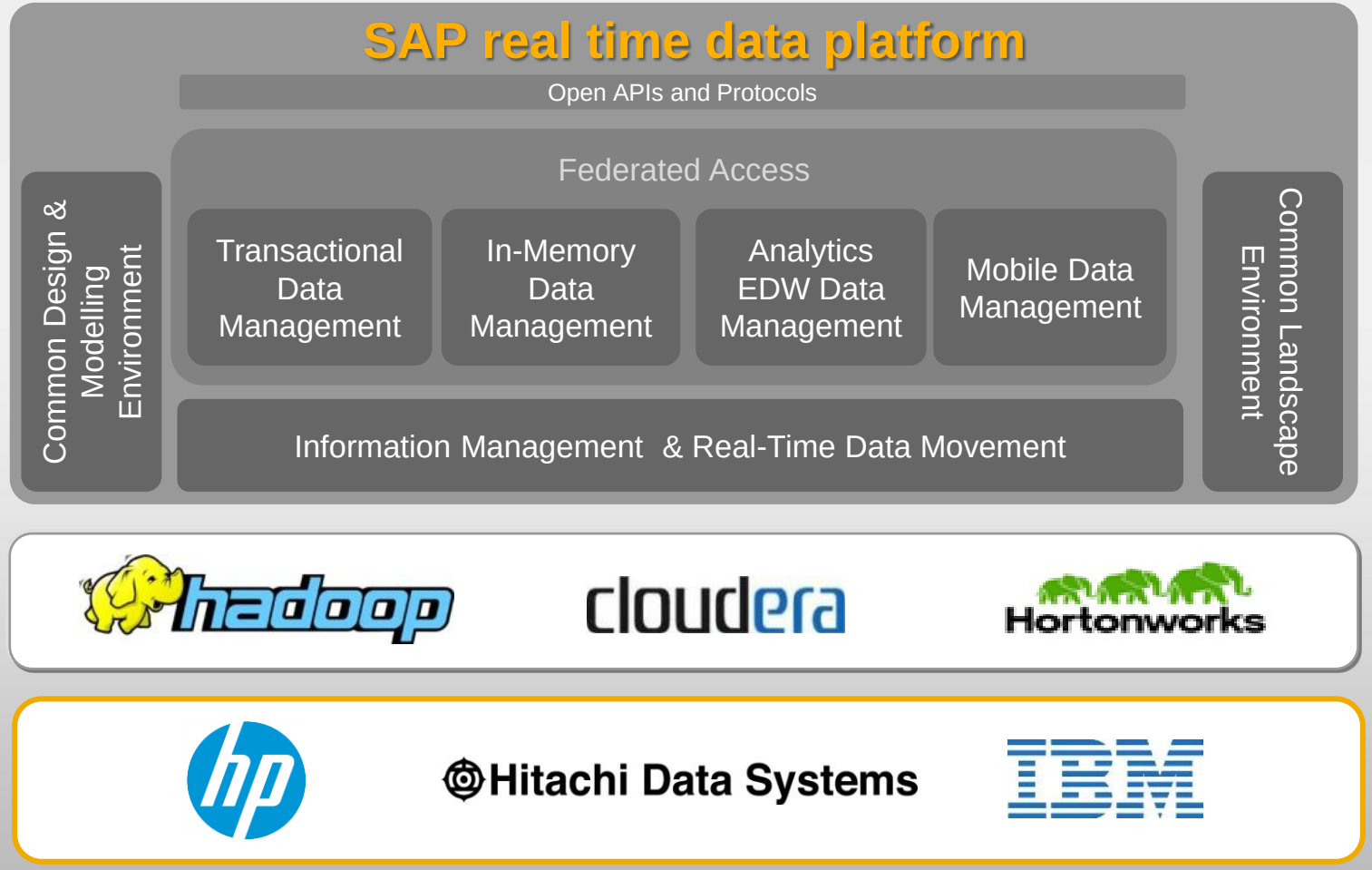
Safe Harbor Statement

The information in this presentation is confidential and proprietary to SAP and may not be disclosed without the permission of SAP. This presentation is not subject to your license agreement or any other service or subscription agreement with SAP. SAP has no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation and SAP's strategy and possible future developments, products and or platforms directions and functionality are all subject to change and may be changed by SAP at any time for any reason without notice. The information on this document is not a commitment, promise or legal obligation to deliver any material, code or functionality. This document is provided without a warranty of any kind, either express or implied, including but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. This document is for informational purposes and may not be incorporated into a contract. SAP assumes no responsibility for errors or omissions in this document, except if such damages were caused by SAP intentionally or grossly negligent.

All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

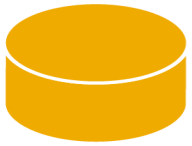
SAP REAL-TIME DATA PLATFORM

A GAME-CHANGER

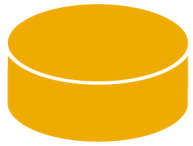


Traditional data management approaches are changing

1980s / 1990s



Today



What is Big Data?

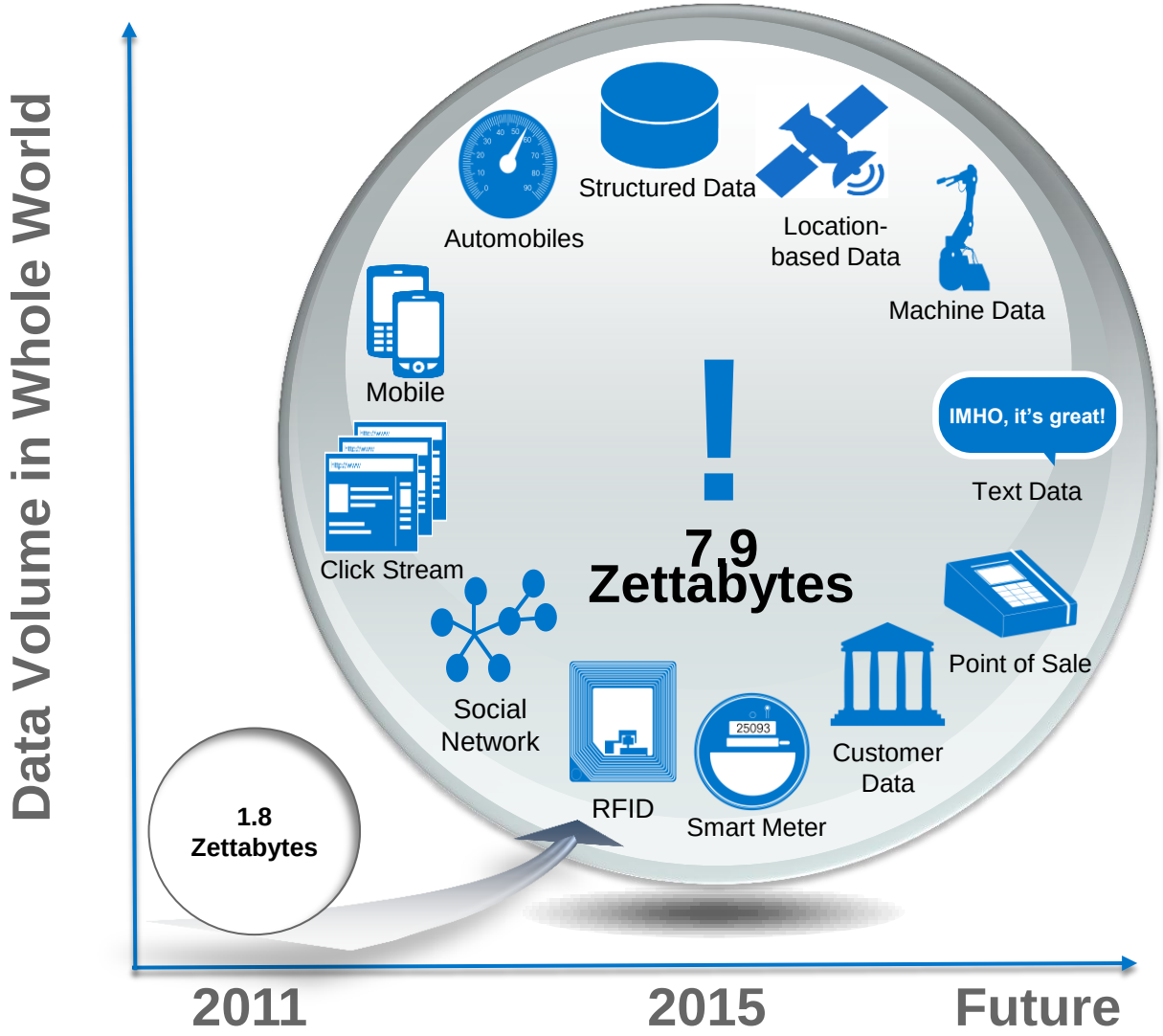
The 3 + 1 V's

Volume
Explosion in the amount of data

Variety
Multiple data formats; non-structured data boom

Velocity
Fast collection, processing and consumption

Value
Keep everything, not only high value data



HANA for Big Data

Key Characteristics: in-memory, row & column, real time

How's HANA for Big Data?

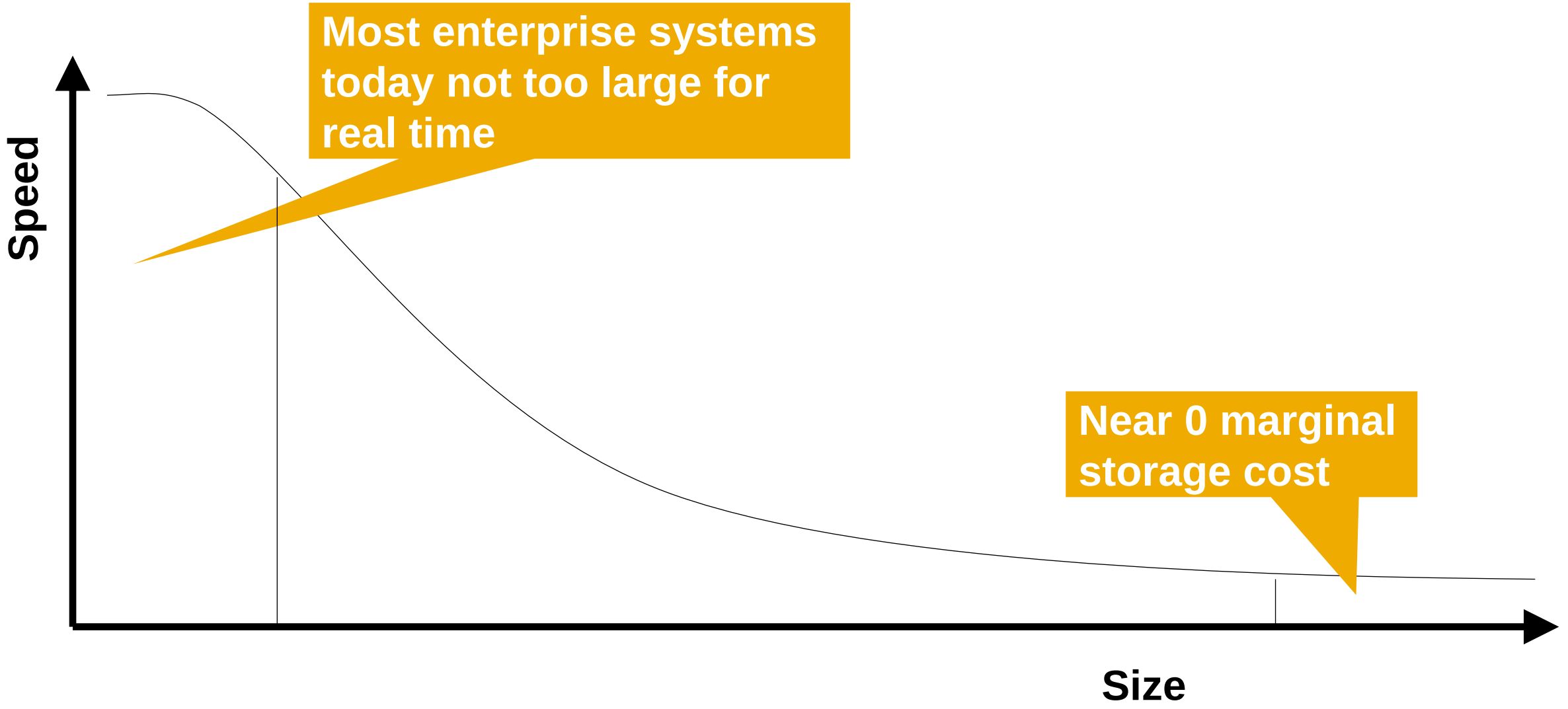
Volume: Billions of records

Variety: Text processing & search

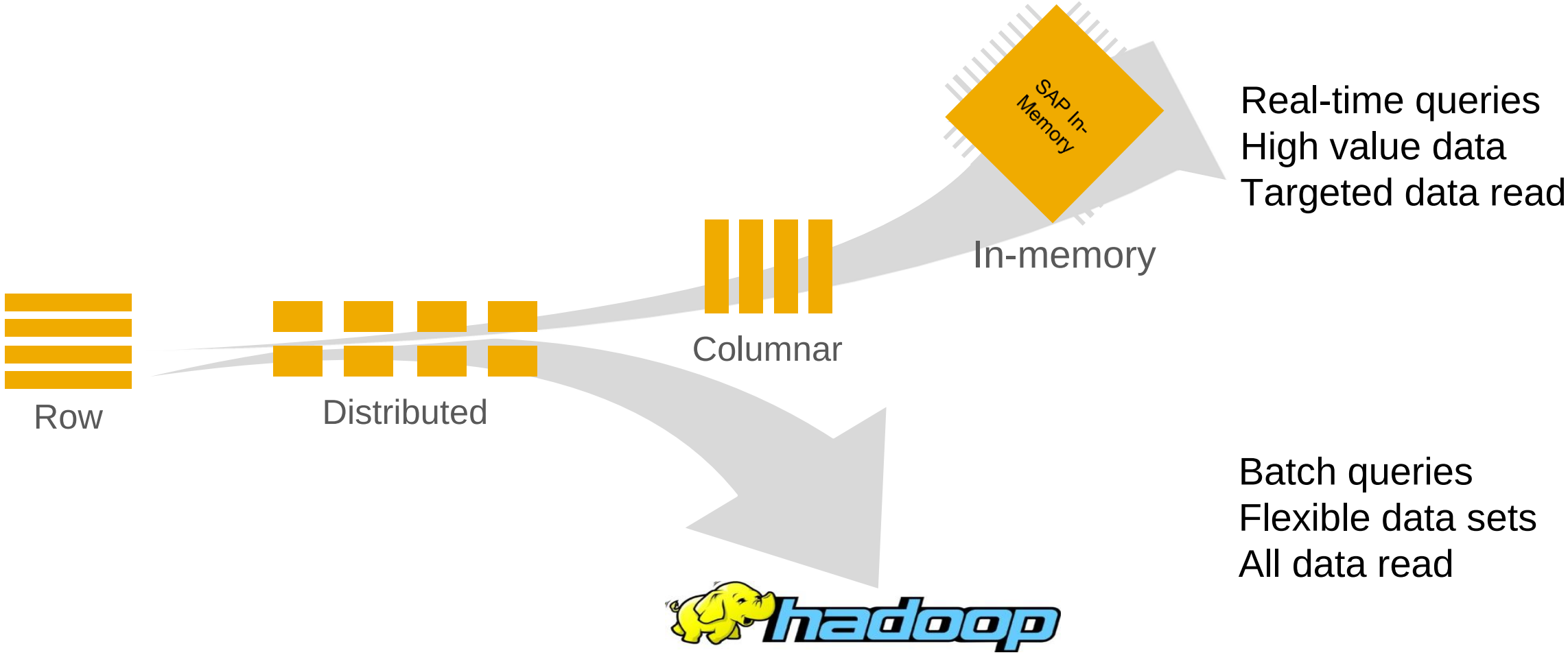
Velocity: Real time

Value: High value data

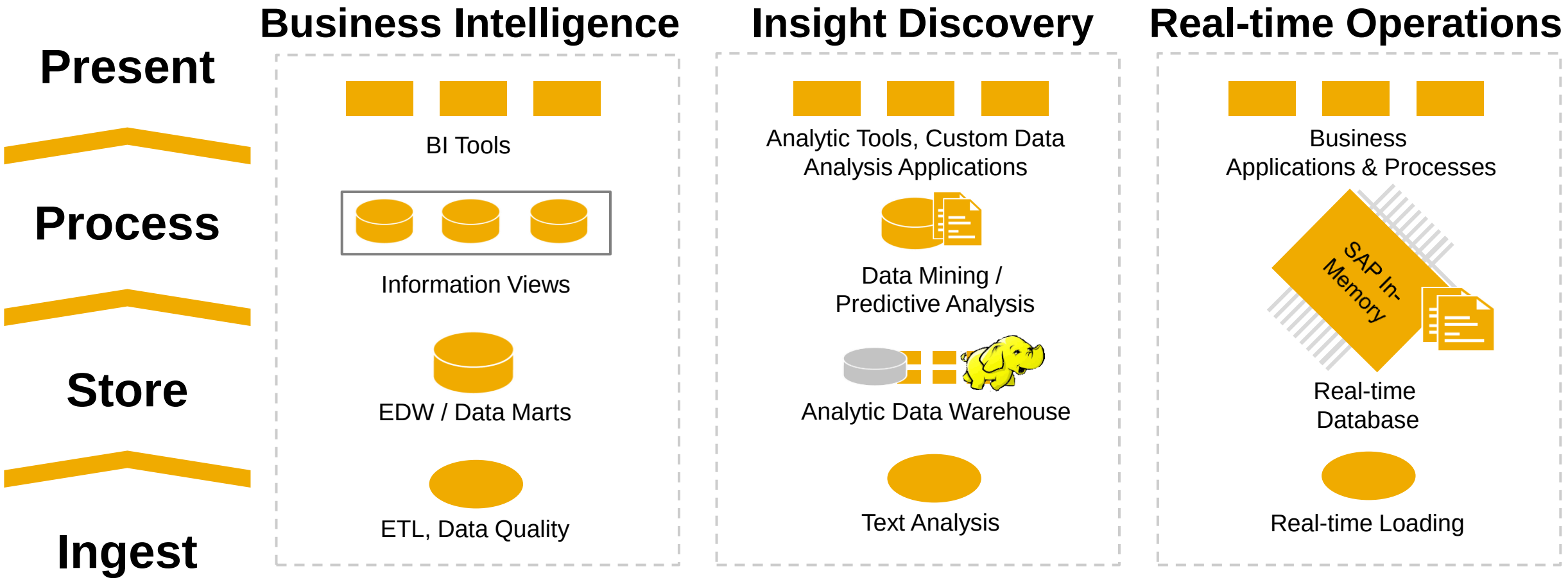
Data management today: systems optimize for speed or capacity



New storage and processing techniques required

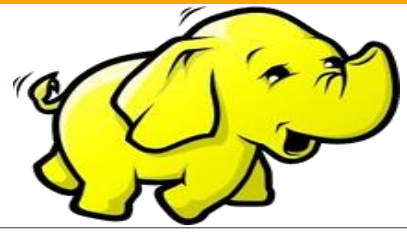


Building an IT landscape for Big Data



Hadoop

What is Apache Hadoop?



Apache Hadoop is **open source** software that enables **reliable**, **scalable**, **distributed** computing on clusters of inexpensive servers

Reliable

- Software is fault tolerant, it expects and handles hardware and software failures

Scalable

- Designed for massive scale of processors, memory, and local attached storage

Distributed

- Handles replication. Offers massively parallel programming model, MapReduce

Hadoop framework handles: partitioning, scheduling, dispatch, execution, communication, failure handling, monitoring, reporting and more

The Apache Hadoop technology family

logical view*

Non-Relational DB

Fine-grained data handling

Hive

"Data warehouse" that provides SQL interface. Data structure is projected ad hoc onto unstructured underlying data

HBase

Column oriented, schema-less, distributed database modeled after Google's BigTable. Random realtime read/write

Scripting

Pig

Platform for manipulating and analyzing large data sets. Scripting language for analysts

Machine Learning

Mahout

Machine learning libraries for recommendations, clustering, classification and itemsets

MapReduce

- Parallel programming
- Large block data handling (e.g. 64MB)

Hadoop Common

HDFS

Distributes & replicates data across machines

MapReduce

Distributes & monitors tasks, restarts failed work

** For simplicity, mappings to servers is omitted*

What does Hadoop bring to the table?

Cost efficient data storage and processing for **large volumes** of structured, semi-structured, and **unstructured data** such as web logs, machine data, text data, call data records (CDRs), audio, video data

Batch Processing

Where fast response times are less critical than reliability and scalability

Complex Information Processing

Enable heavily recursive algorithms, machine learning, & queries that cannot be easily expressed in SQL

Low Value Data Archive

Data stays available, though access is slower

Post-hoc Analysis

Mine raw data that is either schema-less or where schema changes over time

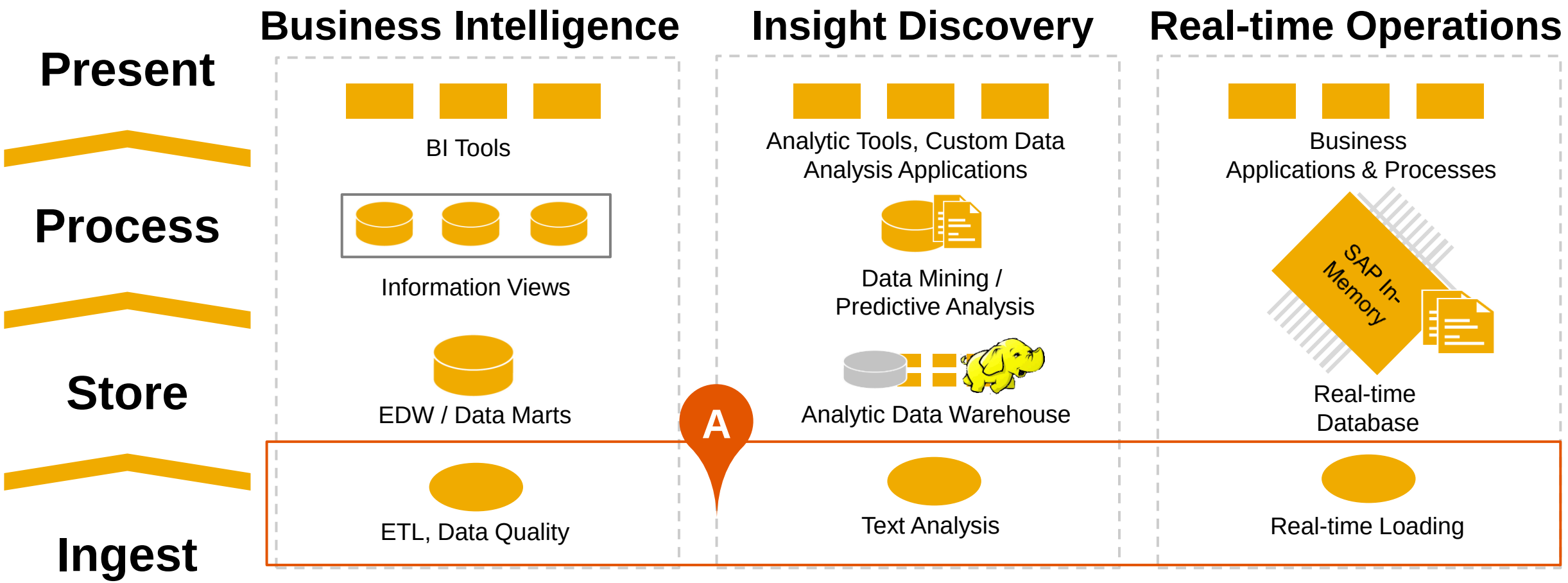
Example: Retail Point of Sales Demo Scenario

http://youtu.be/HmmPje38e1k

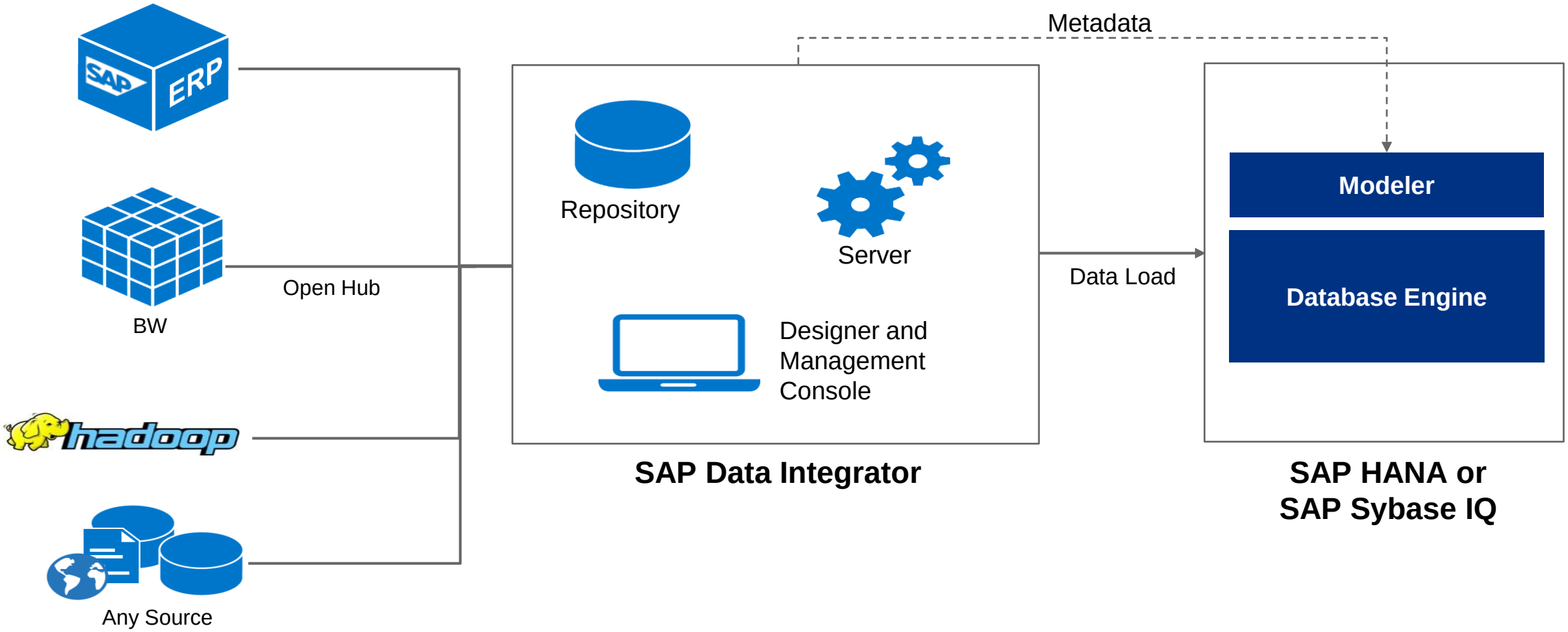
Apache Hadoop bottom line

Strengths	Weaknesses
+ Huge data volumes	- Not efficient at small scale
+ Unstructured data	- Real time is best case challenging, typically not possible
+ Reliable	- Requires skilled engineering, operation and analyst resources
+ Scalable	- Hiring qualified talent
+ Lowest cost	- Less mature than SQL
+ Open source	- Governance
+ No hardware lock in	- Lack of user role support in access model
+ Batch processing	

Hadoop & Enterprise Information Management

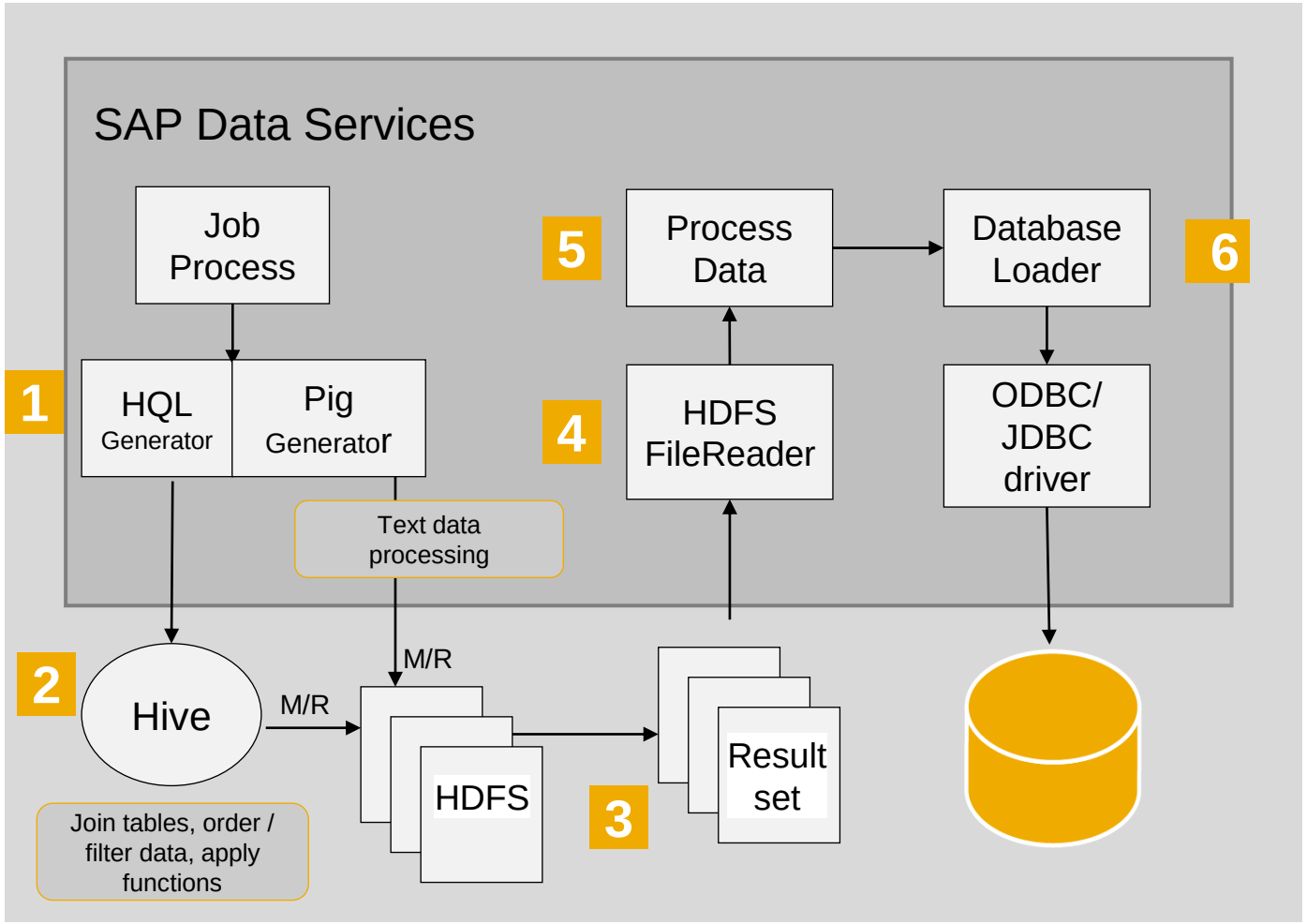


IT administrator: Extract, transform, and load data quickly

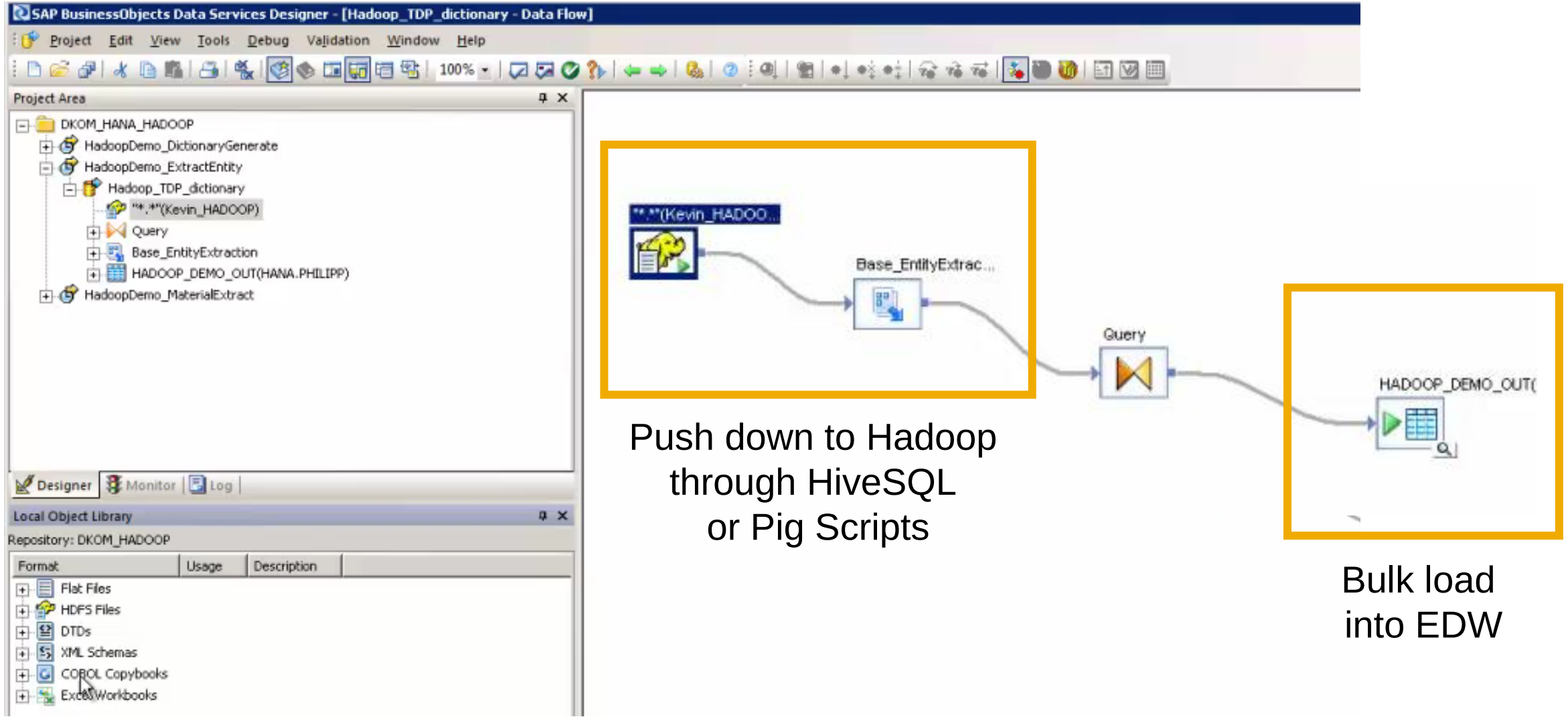


Loading data from Hadoop into your database

1. Based on target, SAP Data Services translates queries into:
   - Hive Query Language (HQL) → Hive
   - Pig script → HDFS
2. Hive/Pig converts queries to Map/Reduce jobs
3. Result data files are generated on the HDFS system
4. SAP Data Services use multiple threads to process data from Hive/Pig
5. Optional transforms: Data quality operations
6. Load results into database



SAP Data Services: Simple GUI build and run ETL process



Processing text to extract relevant data from Hadoop

1 Use SAP Data Services to **extract**:

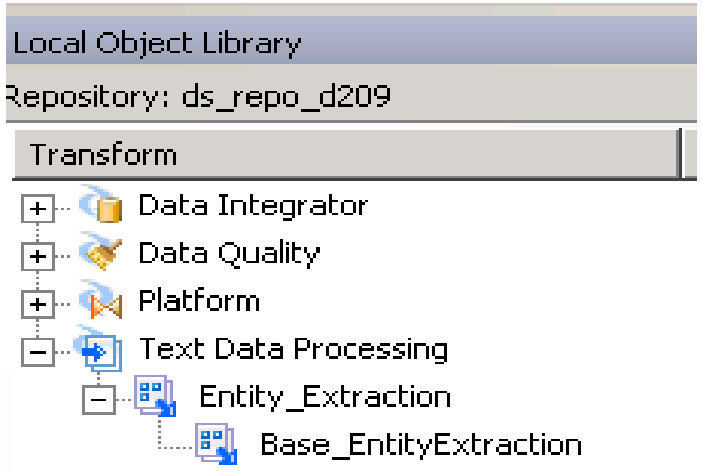
- Core entities (who, what, when, where, etc.)
- Domains (voice of customer, public sector, enterprise, etc)
- Sentiment analysis (strong positive, weak positive, neutral, weak negative, strong negative)

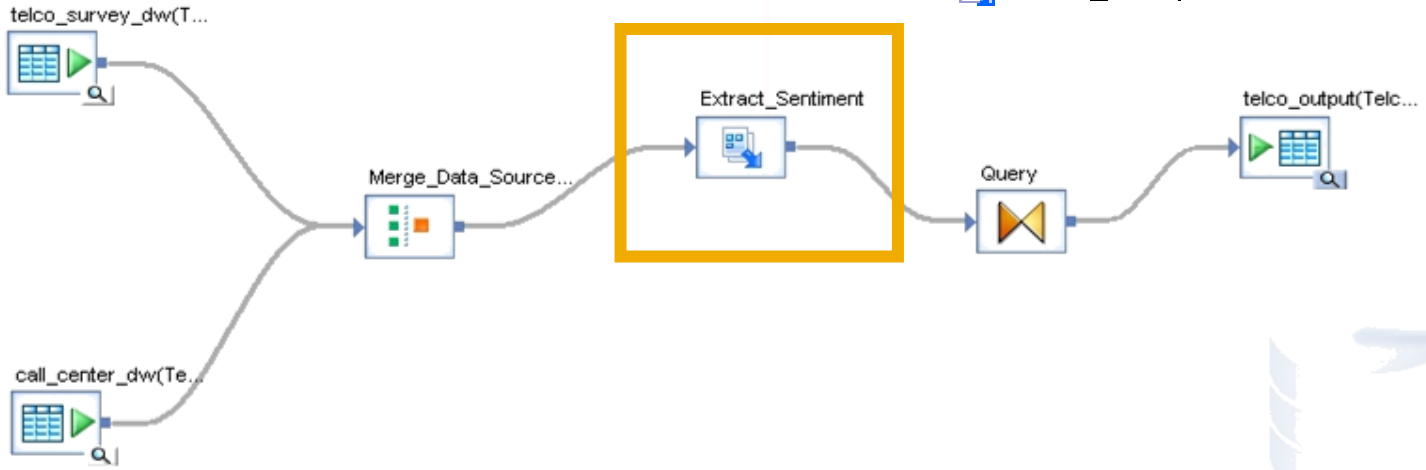
2 Perform **transformations**

- Map text into pre-defined structures
- Cleanse, match, de-duplicate data

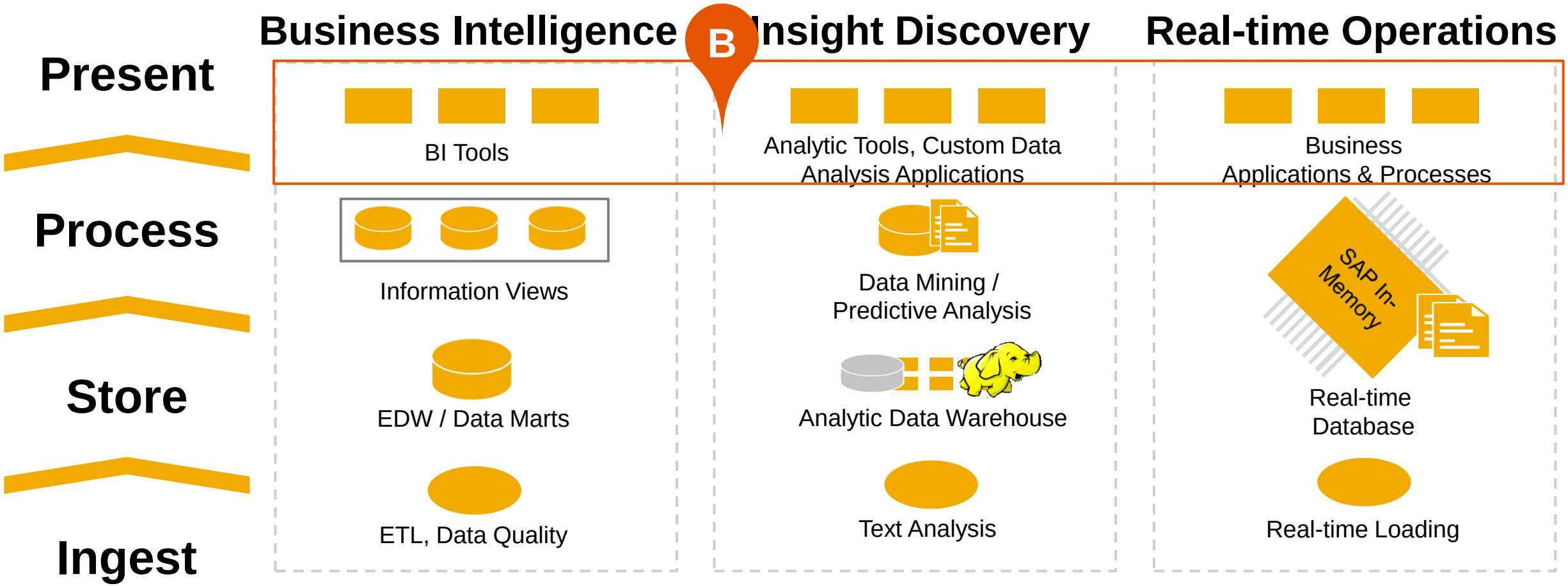
3 **Load** results quickly into EDW

- Map text to structure

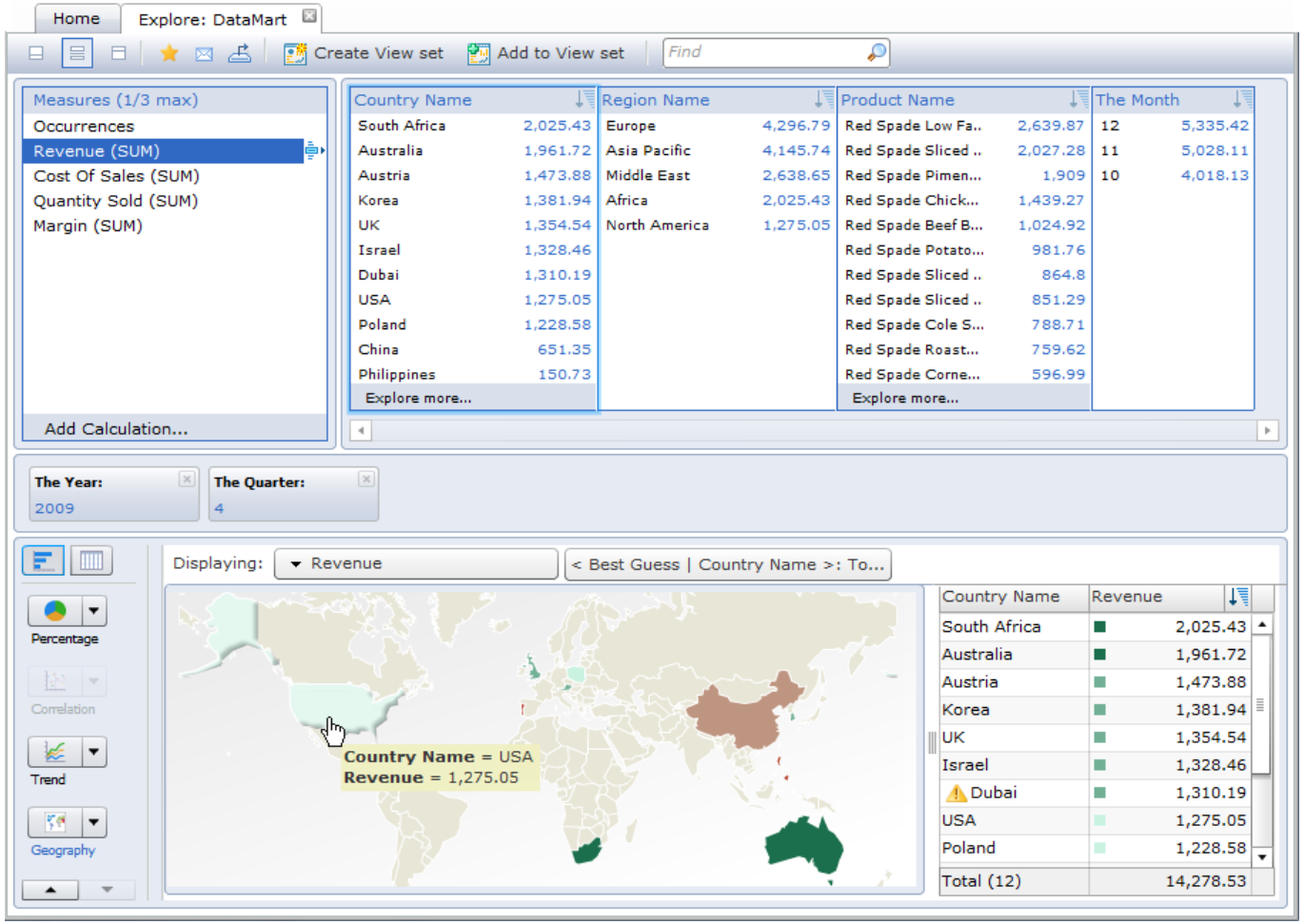


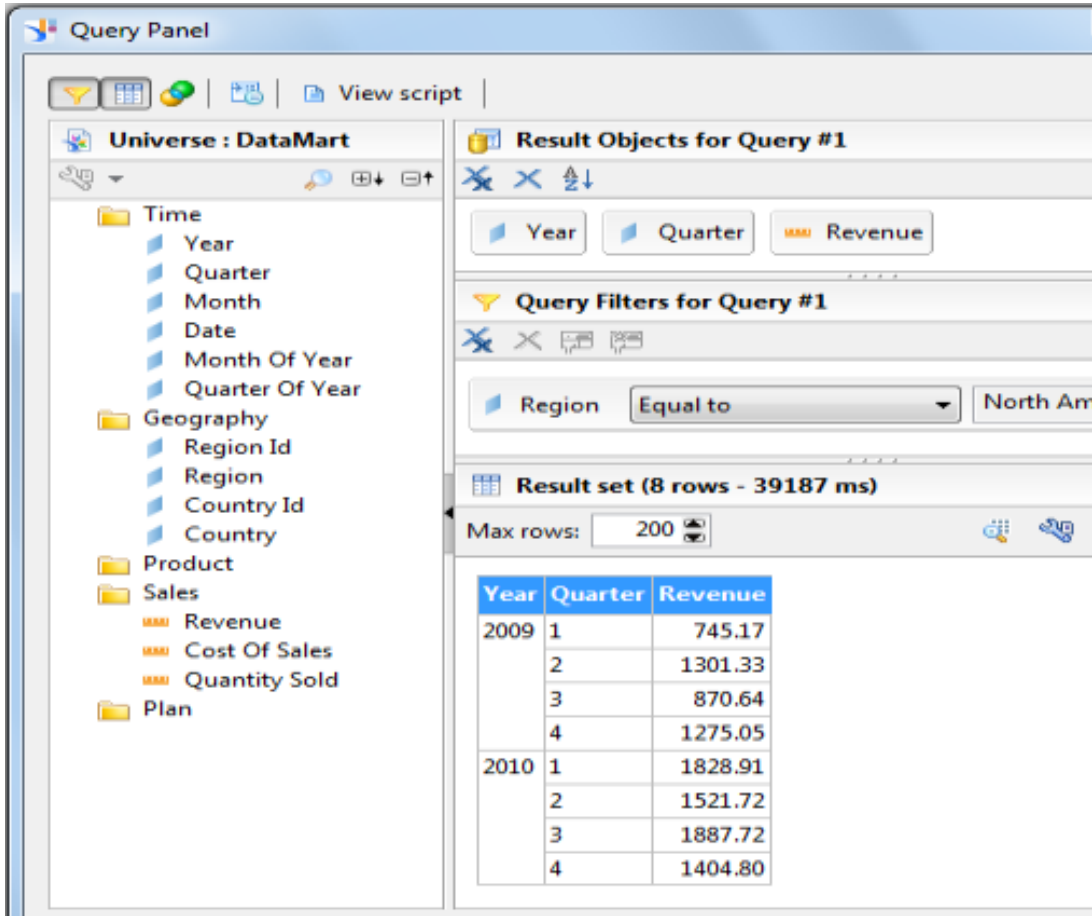


Hadoop Analytics



Business Analyst: Viewing data in Hadoop using GUI tools



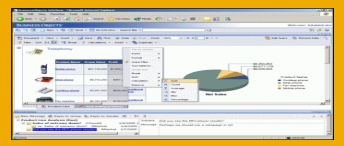


Automatically generates HiveQL statements that are executed on a Hadoop cluster

SAP BusinessObjects BI: Hadoop for Business Analysts

Empower all analysts, enable all workflows

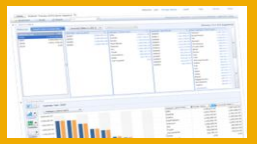
Common user experience for all front-end tools



Web Intelligence

Crystal Reports

Dashboards



Explorer

High performance, feature rich, secure

Best access method for each specific data source

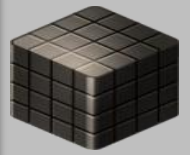
Universe Access

Direct Access

Extract, define, & manipulate metadata

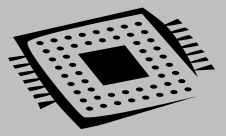
HADOOP HIVE

All data sources

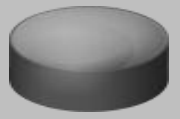


SAP BW

Sybase databases



SAP HANA



3rd party databases

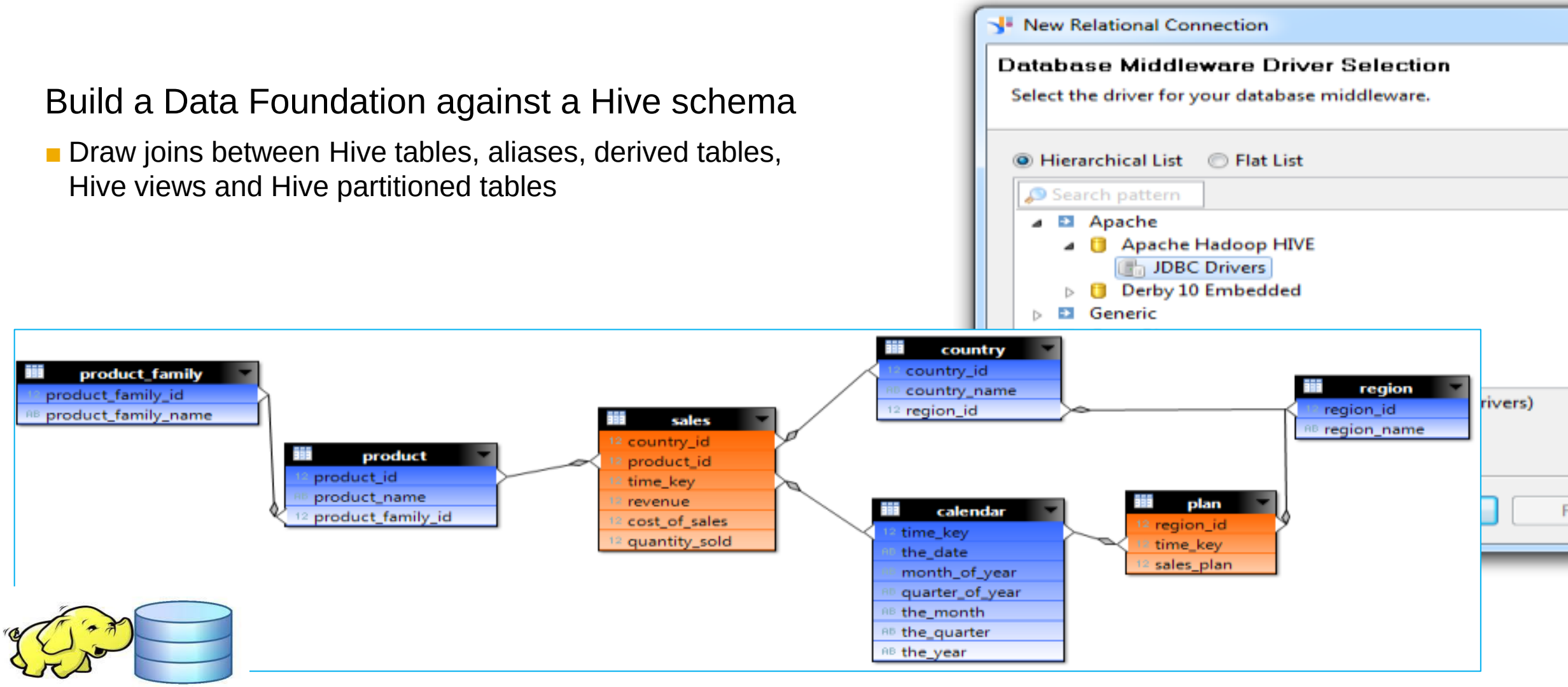
Files

Web Services

Simple Tools for the BI administrator to define data access

Build a Data Foundation against a Hive schema

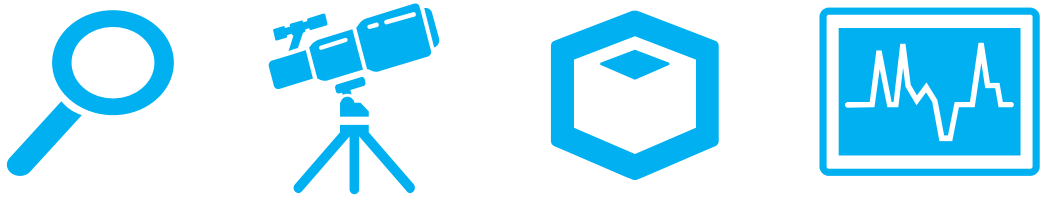
- Draw joins between Hive tables, aliases, derived tables, Hive views and Hive partitioned tables



Data Scientist: Flexibility is of the essence

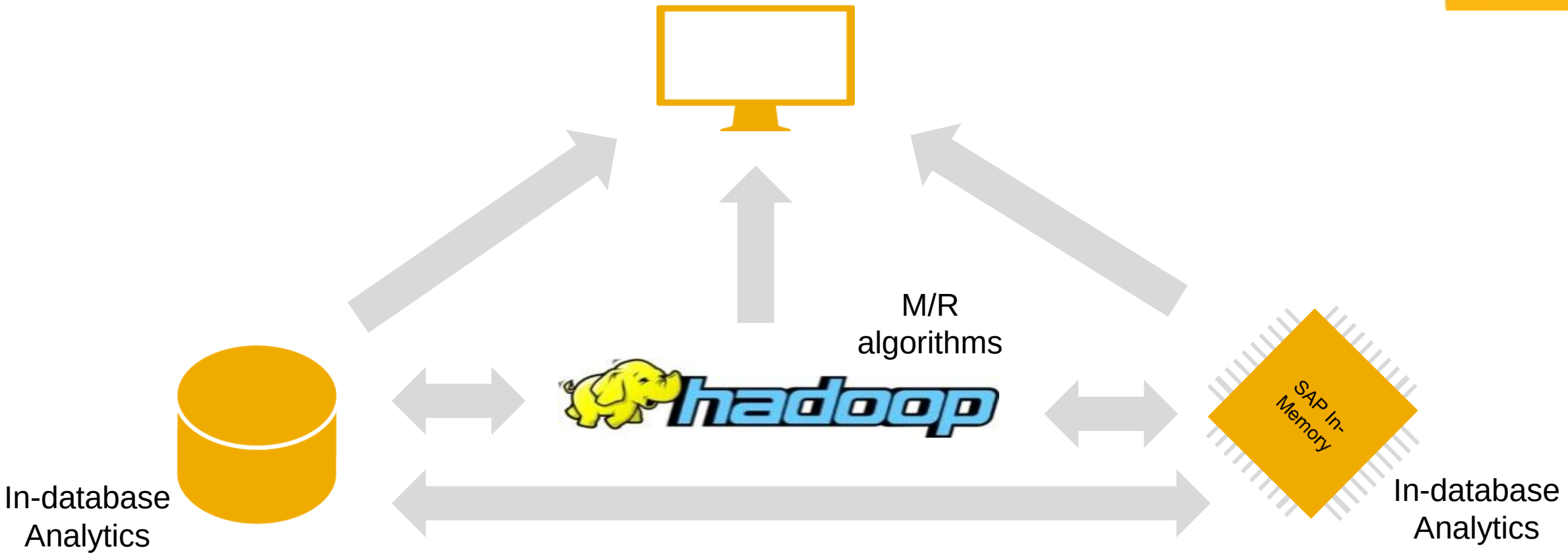
Chooses the variables that offer the most promise

Chooses the best tool based on data mining technique



Chooses best analysis engine based on algorithm & data

Data Scientist: Open integration is key



Building an IT landscape for Big Data

Engage

Process

Store

Ingest

Business Intelligence

BI Tools

Information Views

EDW / Data Marts

ETL, Data Quality

Insight Discovery

Analytic Tools, Custom Data Analysis Applications

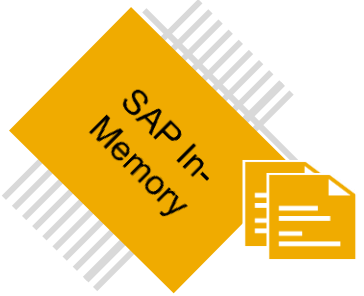
Data Mining / Predictive Analysis

Analytic Data Warehouse

Text Analysis

Real-time Operations

Business Applications & Processes



Real-time Database

Real-time Loading

Questions?

william.gardella@sap.com

© 2012 SAP AG. All rights reserved.

© 2012 SAP AG. Alle Rechte vorbehalten.