

HANA & Hadoop for Big Data Management



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



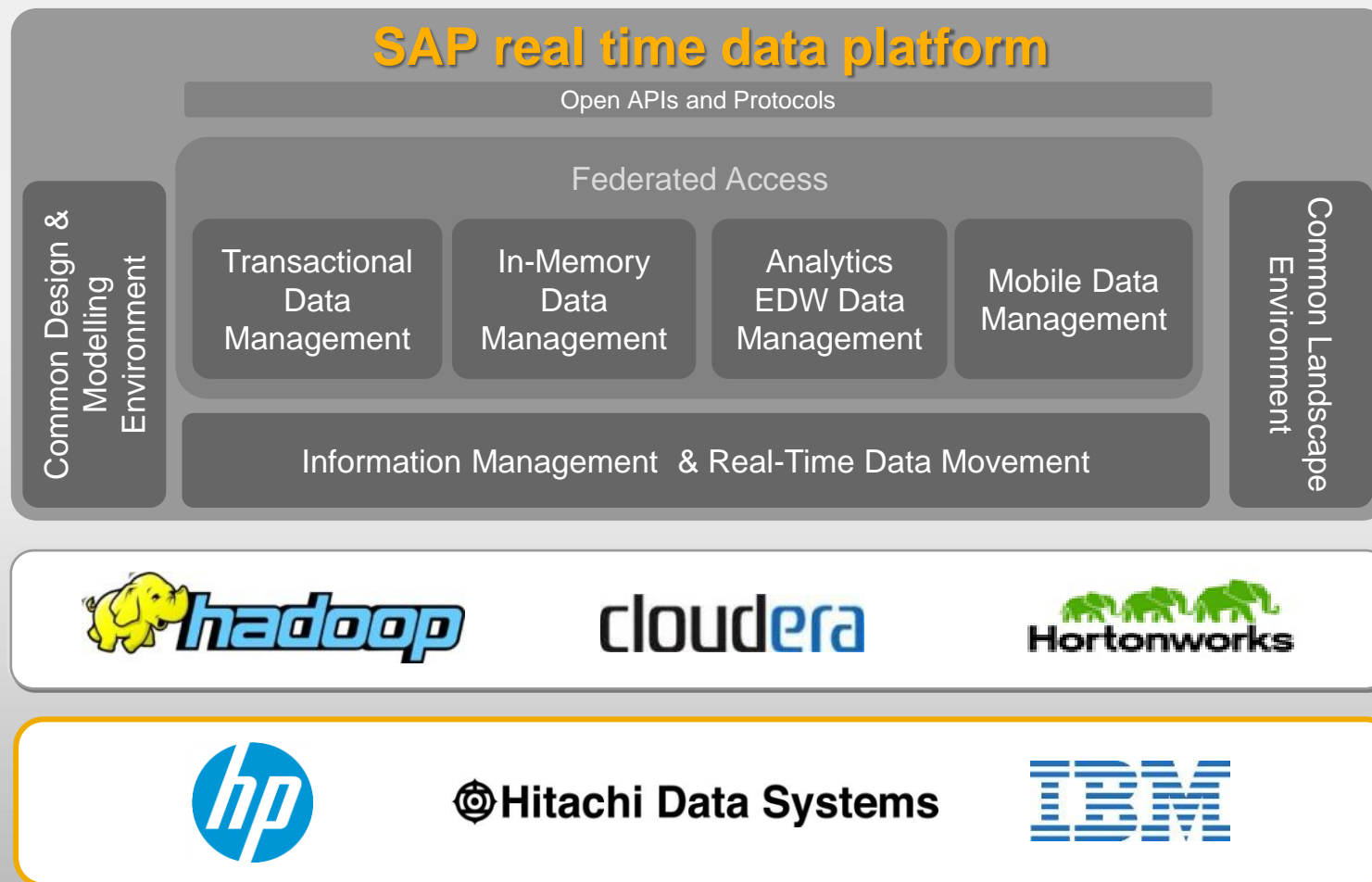
Safe Harbor Statement

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.

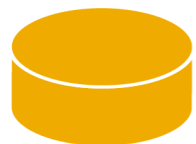
SAP REAL-TIME DATA PLATFORM

A GAME-CHANGER

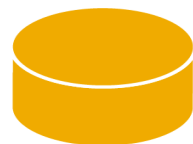


Traditional data management approaches are changing

1980s / 1990s



Today



100101
011010
100101



What is Big Data?

The 3 + 1 V's

Volume

Explosion in the amount of data

Variety

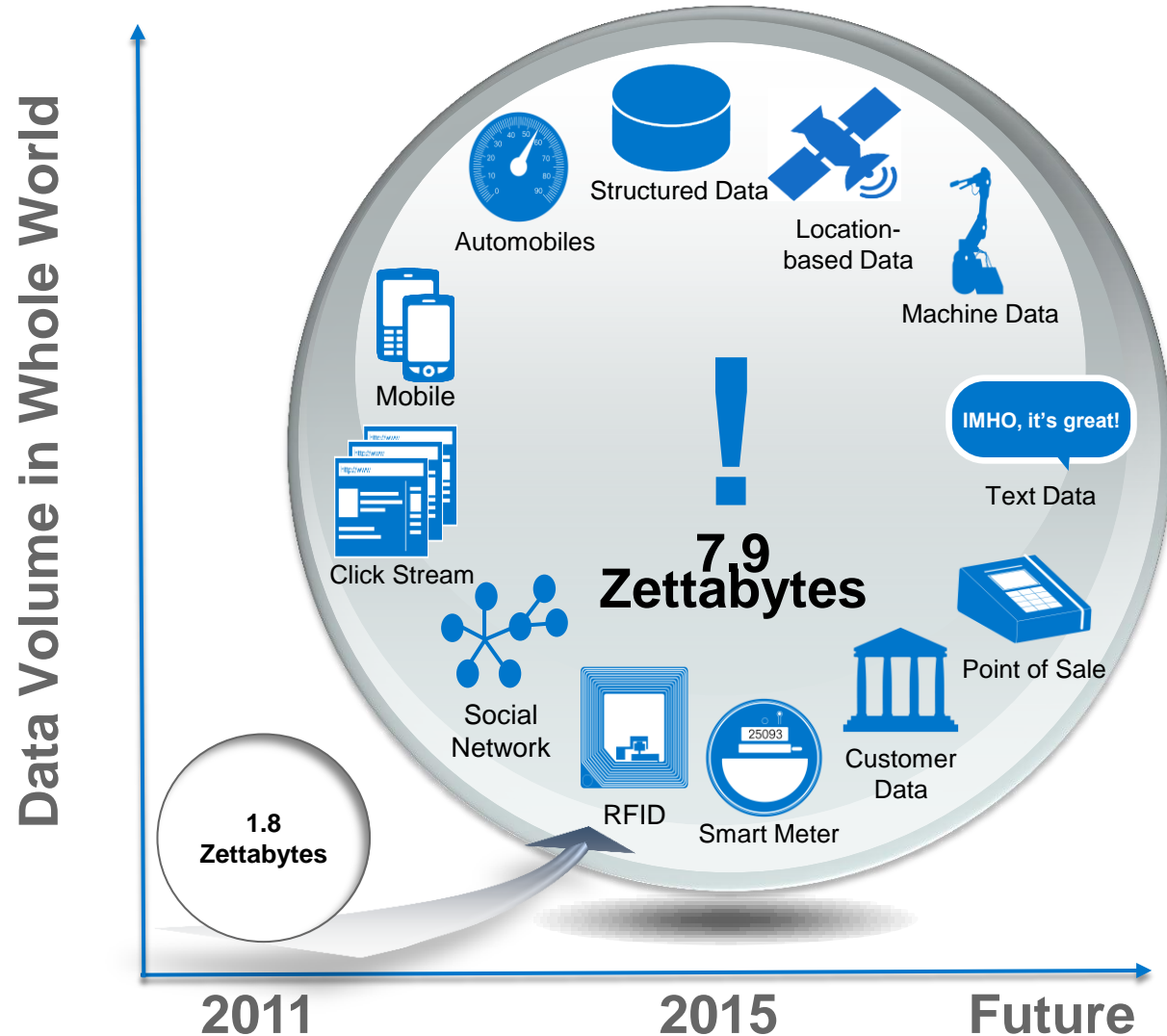
Multiple data formats;
non-structured data boom

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

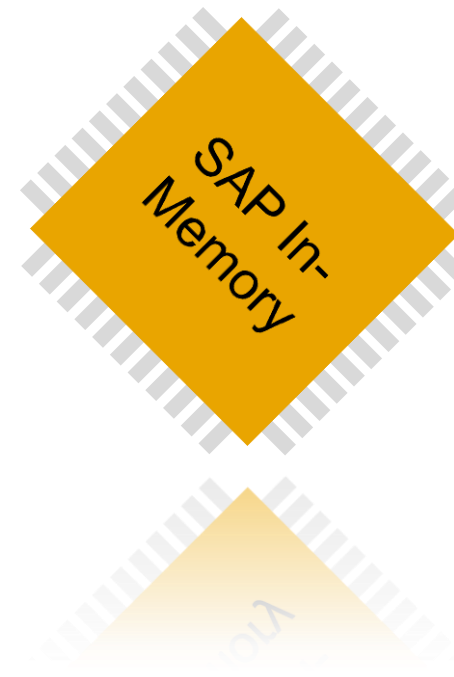
How's HANA for Big Data?

Volume: Billions of records

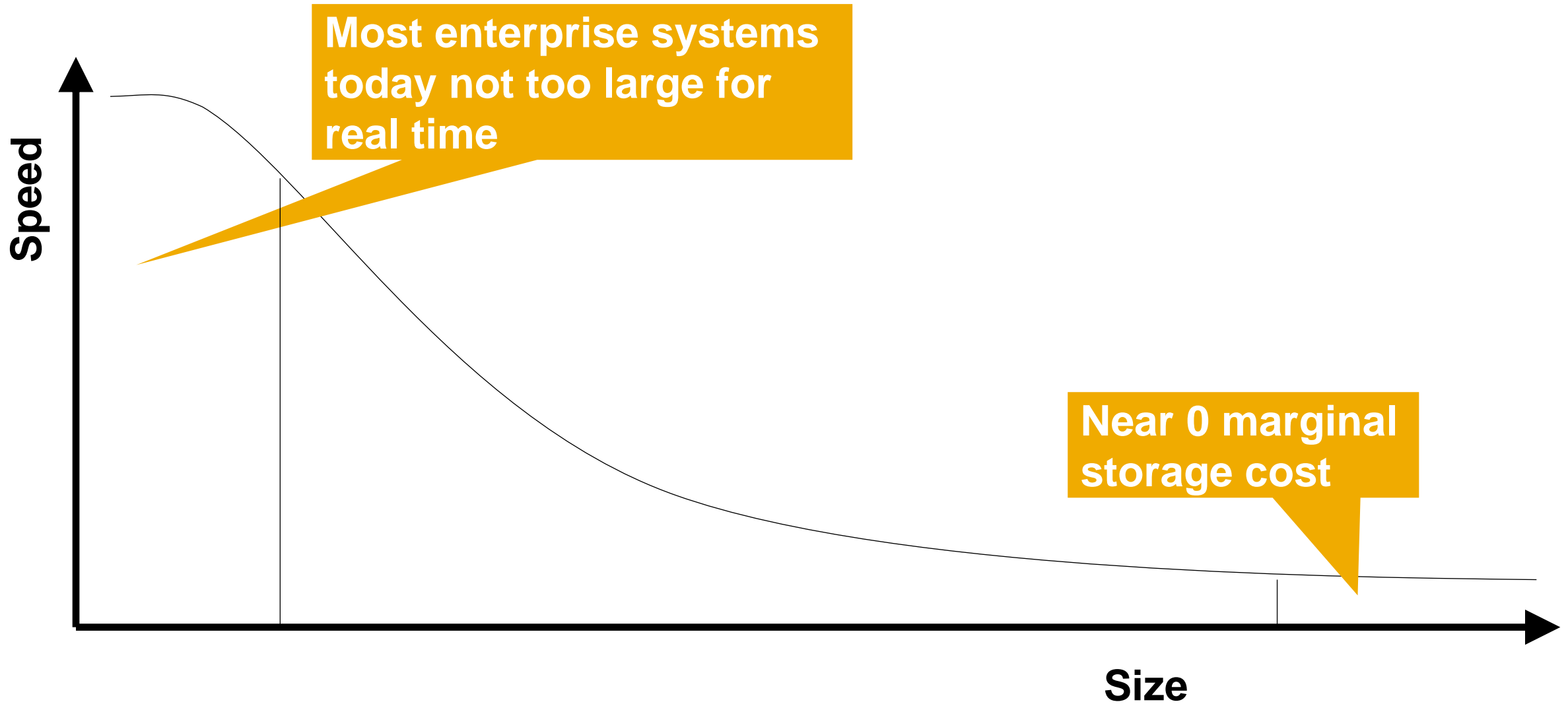
Variety: Text processing & search

Velocity: Real time

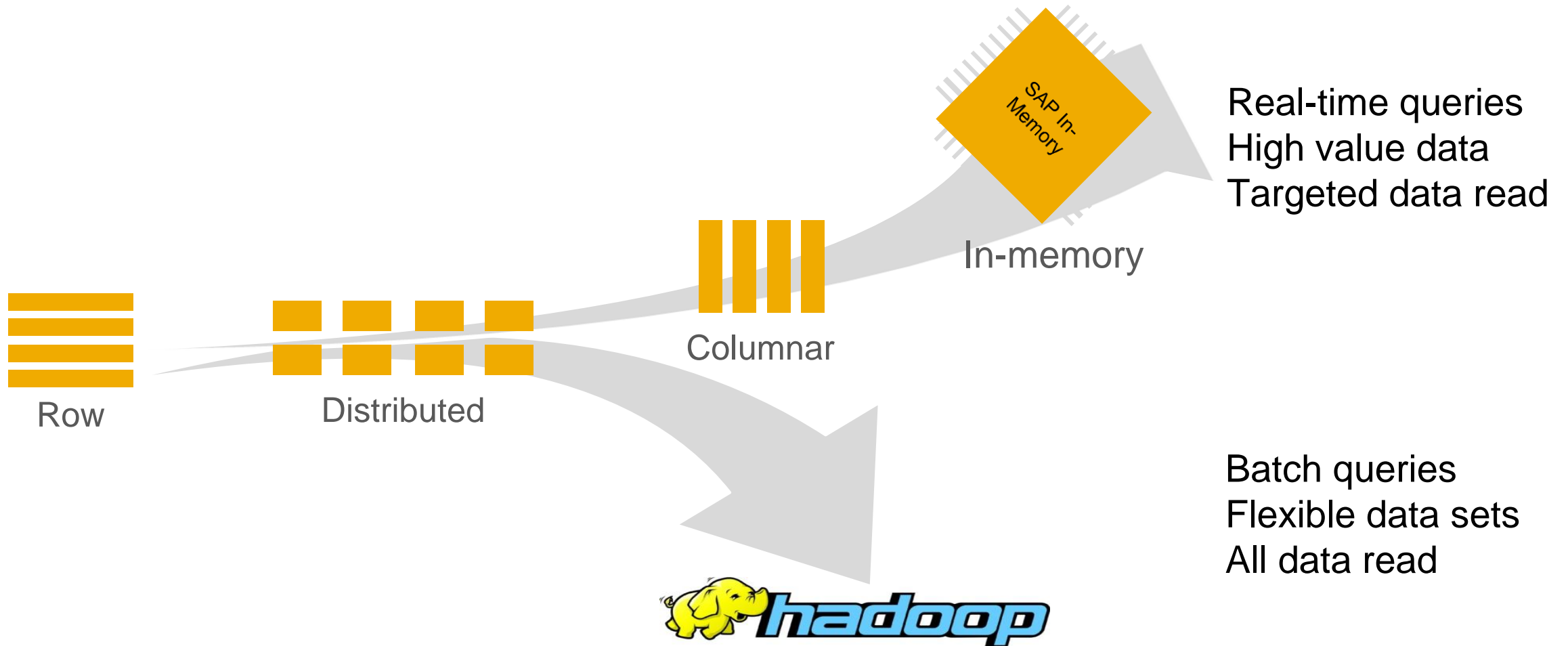
Value: High value data



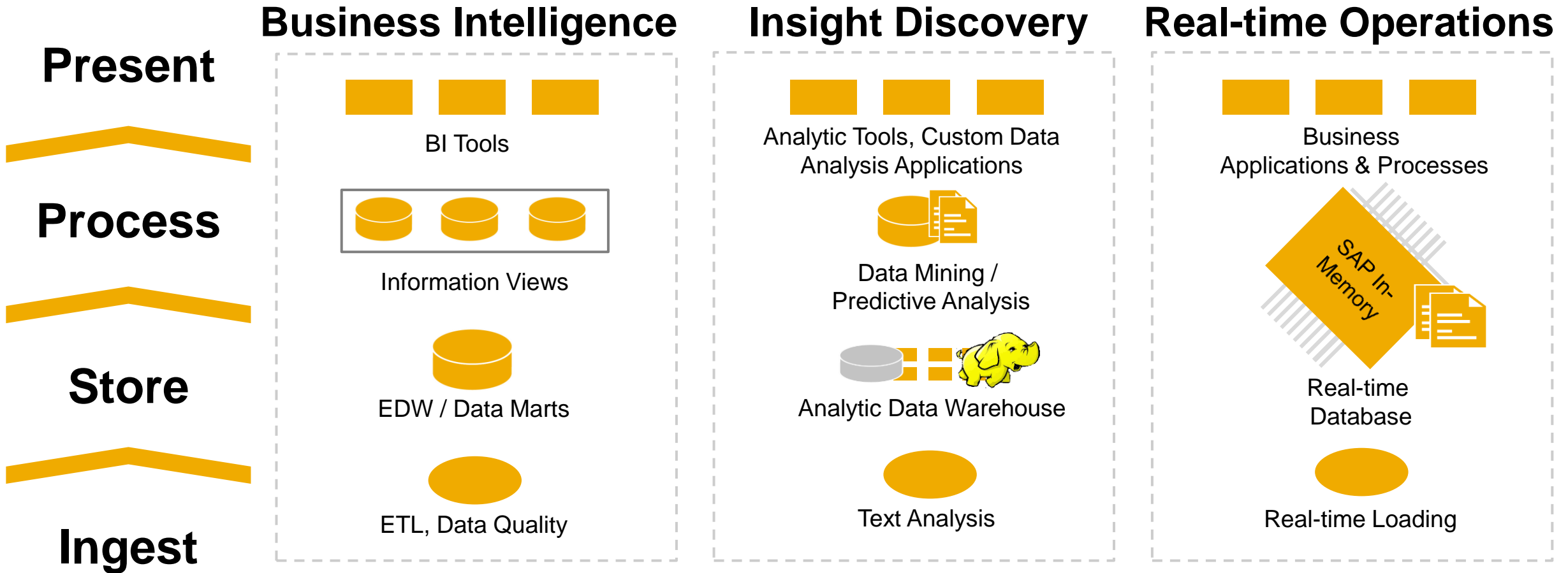
Data management today: systems optimize for speed or capacity



New storage and processing techniques required



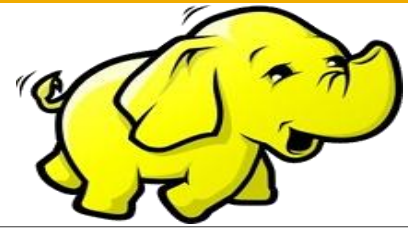
Building an IT landscape for Big Data





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

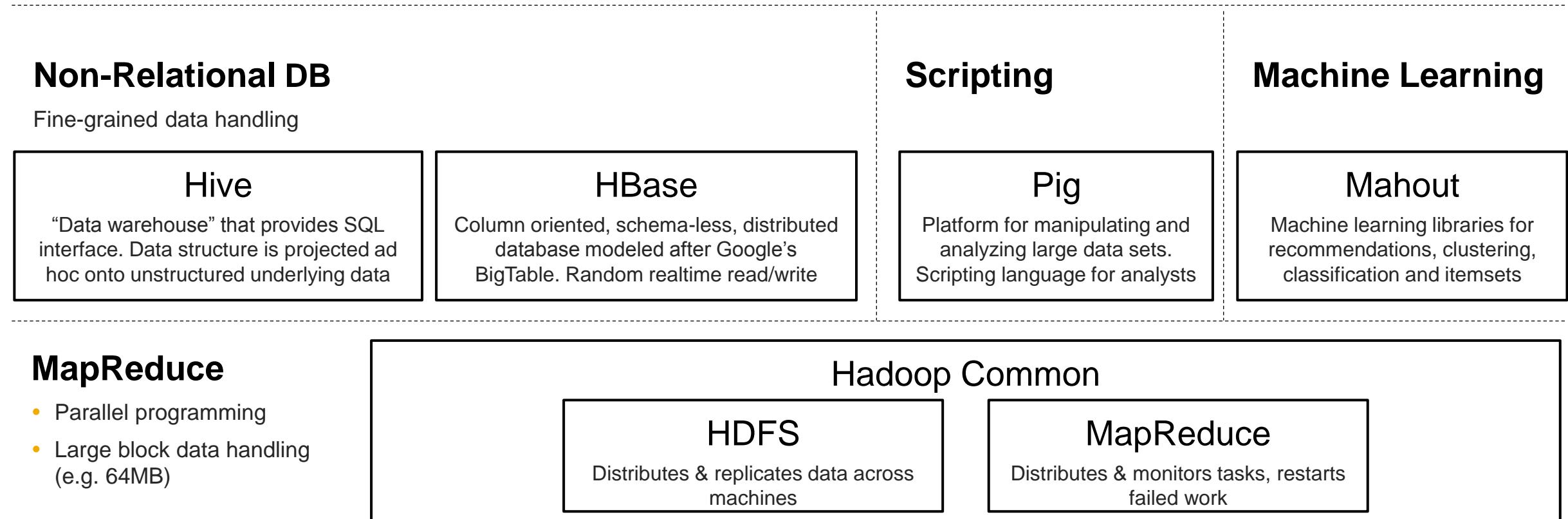
Distributed

- 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*



* 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

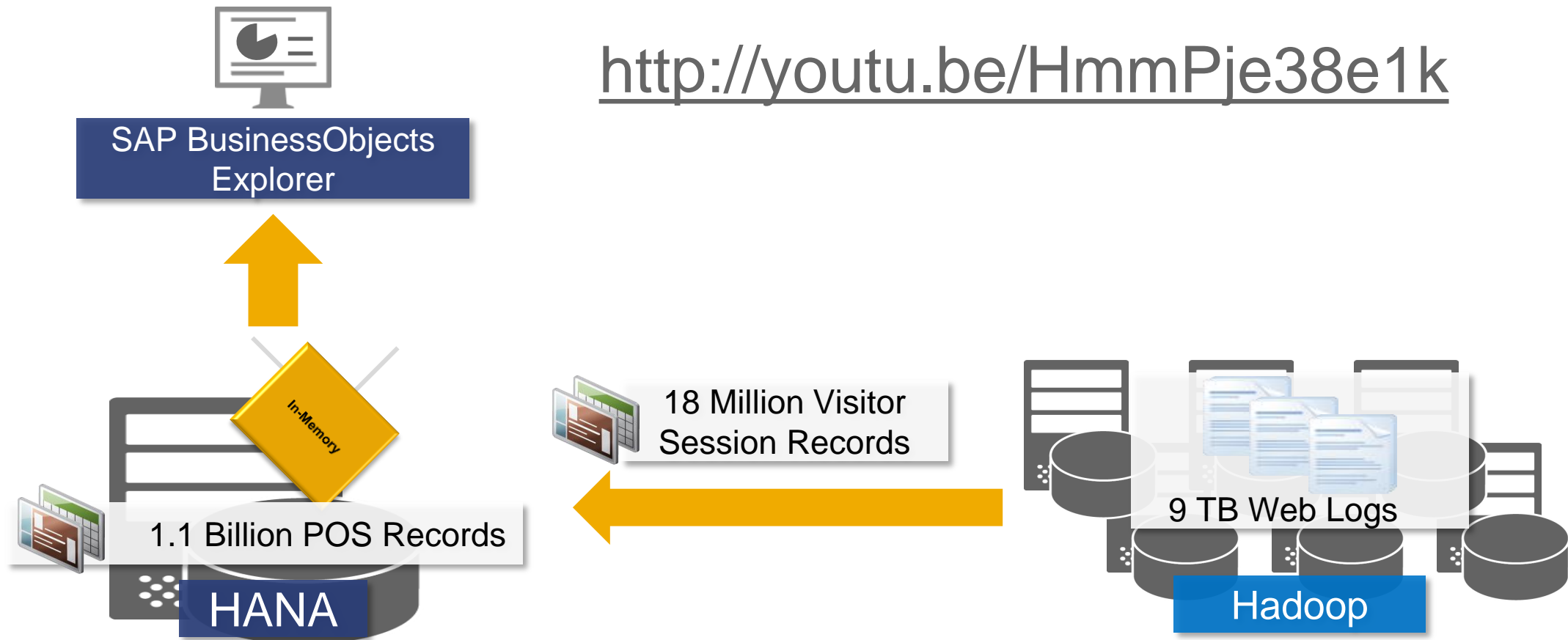
Low Value Data Archive

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



Apache Hadoop bottom line

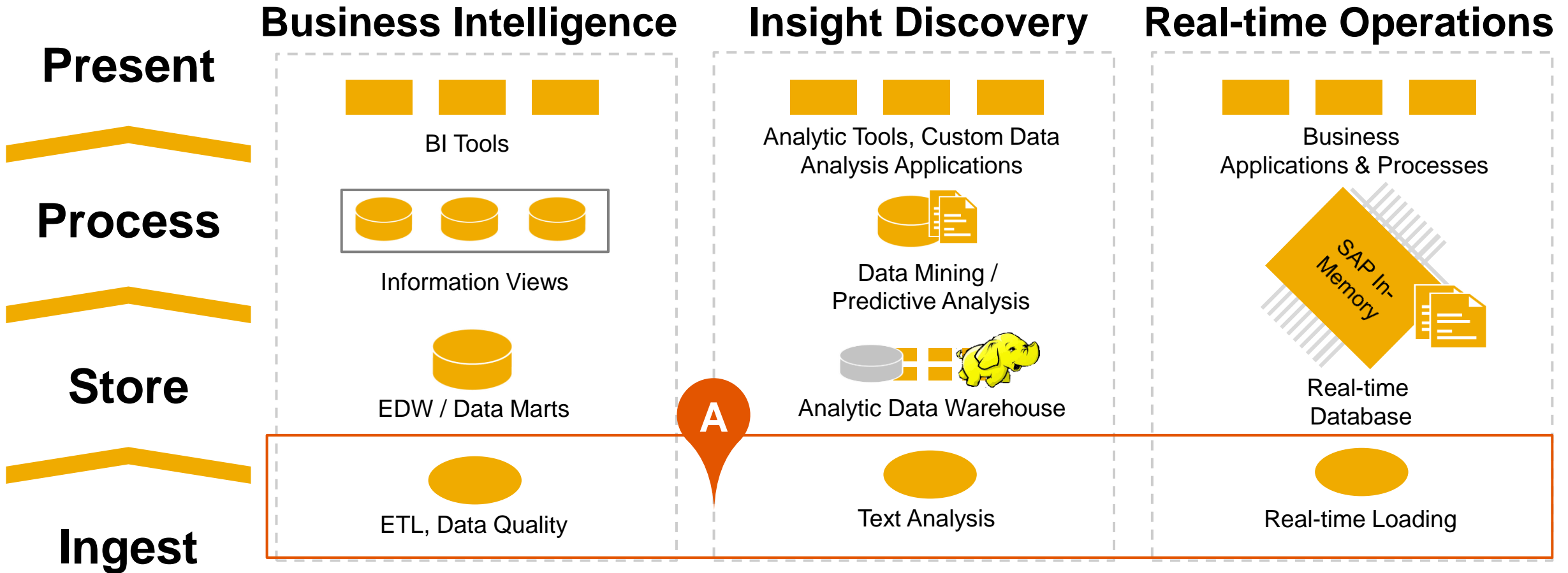
Strengths

- + Huge data volumes
- + Unstructured data
- + Reliable
- + Scalable
- + Lowest cost
- + Open source
- + No hardware lock in
- + Batch processing

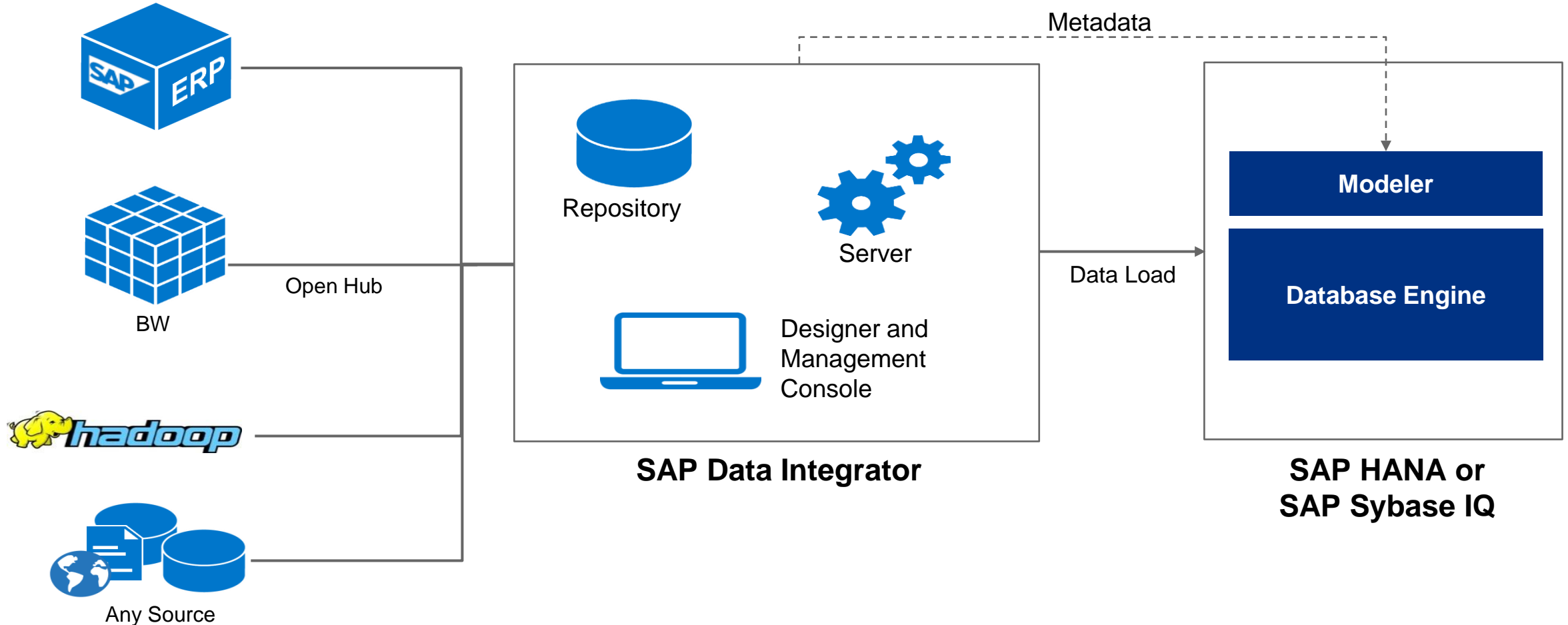
Weaknesses

- Not efficient at small scale
- Real time is best case challenging, typically not possible
- Requires skilled engineering, operation and analyst resources
- Hiring qualified talent
- Less mature than SQL
- Governance
- Lack of user role support in access model

Hadoop & Enterprise Information Management

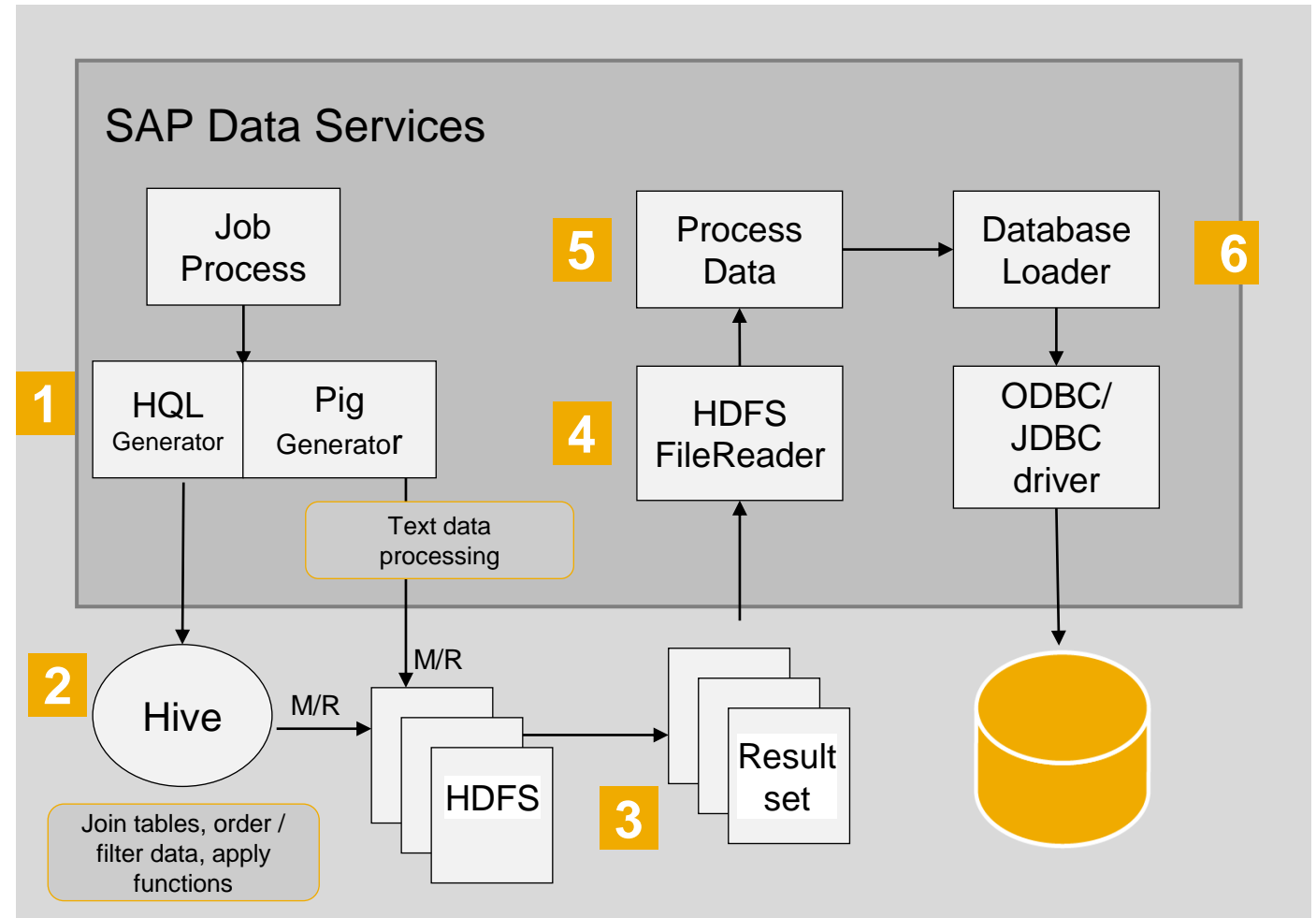


IT administrator: Extract, transform, and load data quickly

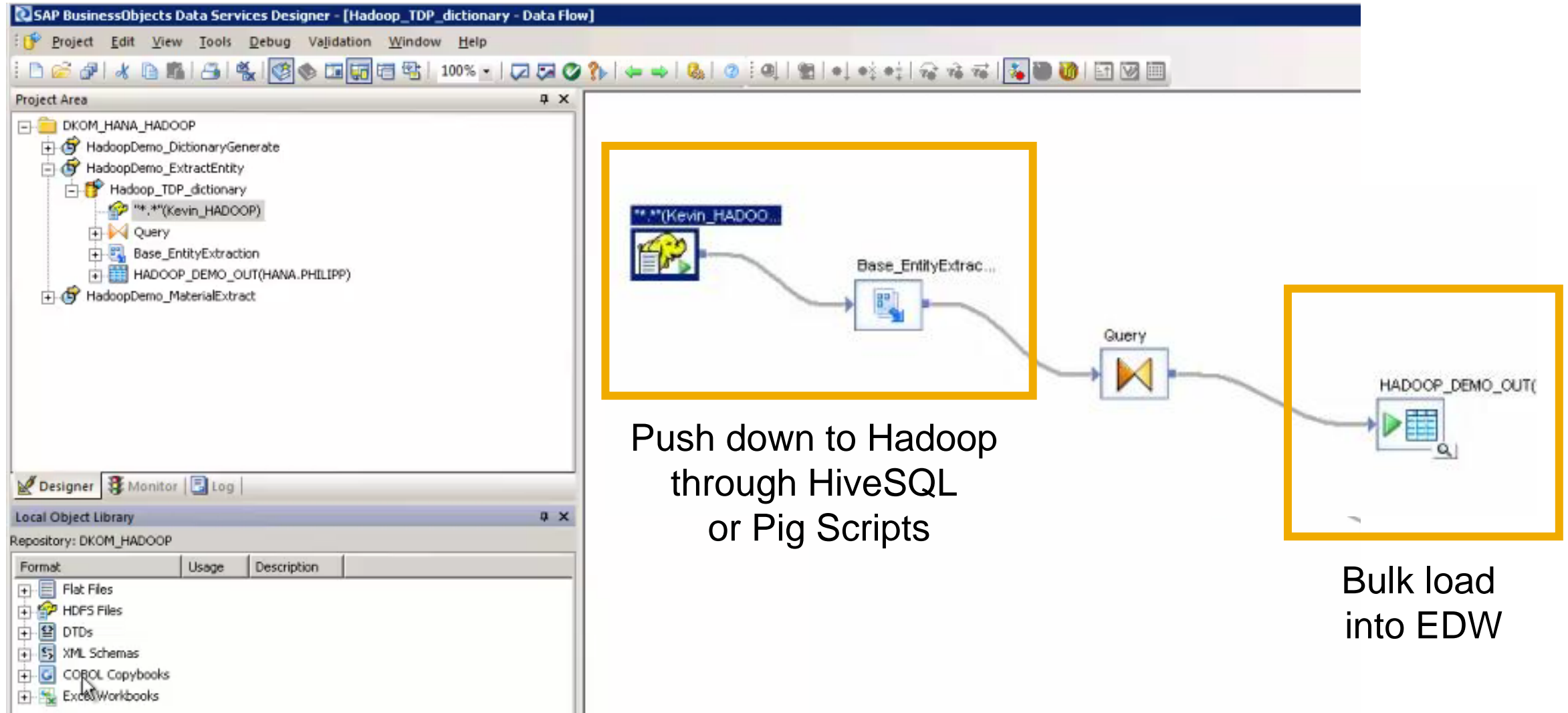


Loading data from Hadoop into your database

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**:

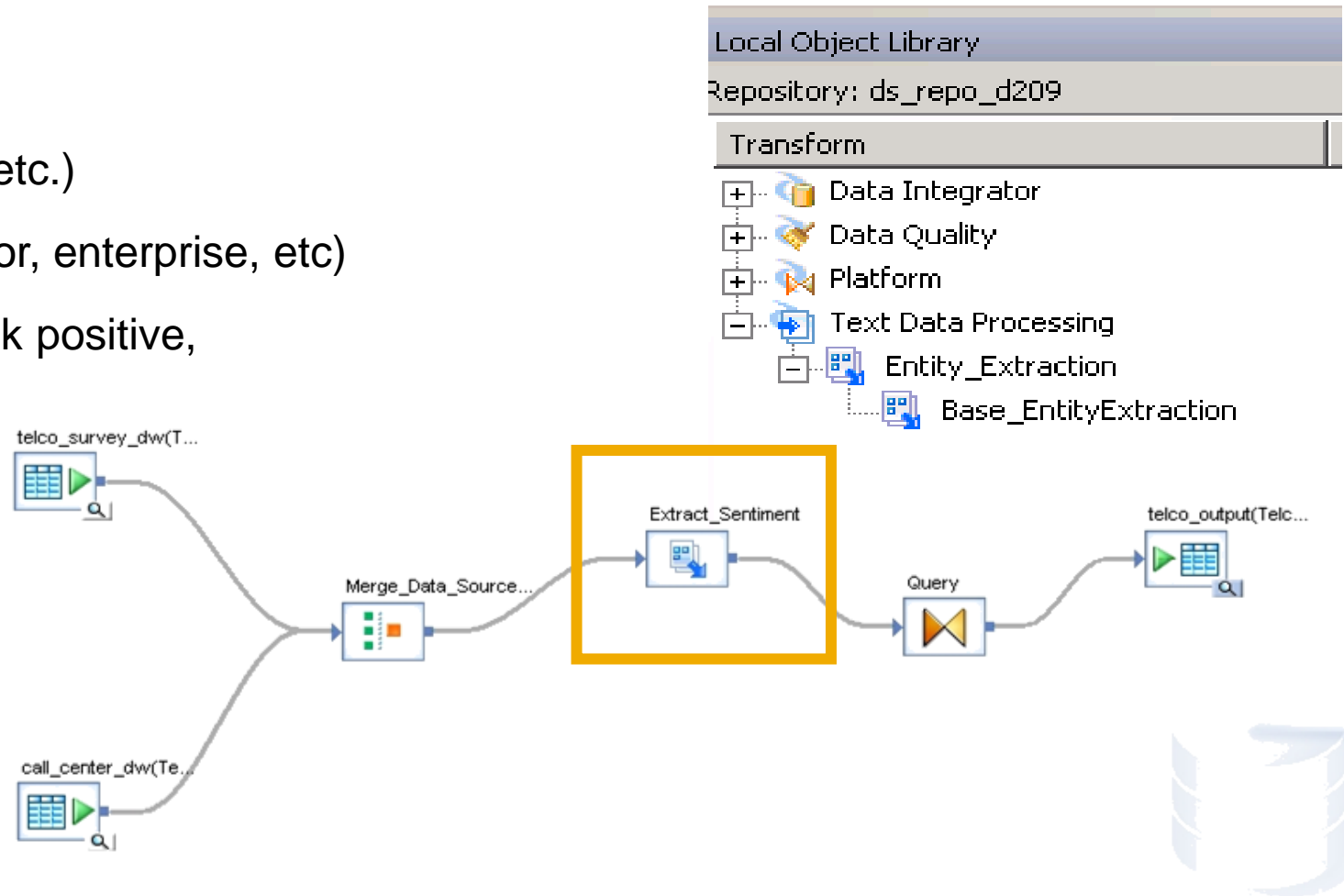
- ❑ Core entities (who, what, when, where, etc.)
- ❑ Domains (voice of customer, public sector, enterprise, etc)
- ❑ Sentiment analysis (strong positive, weak positive, neutral, weak negative, strong negative)

2 Perform **transformations**

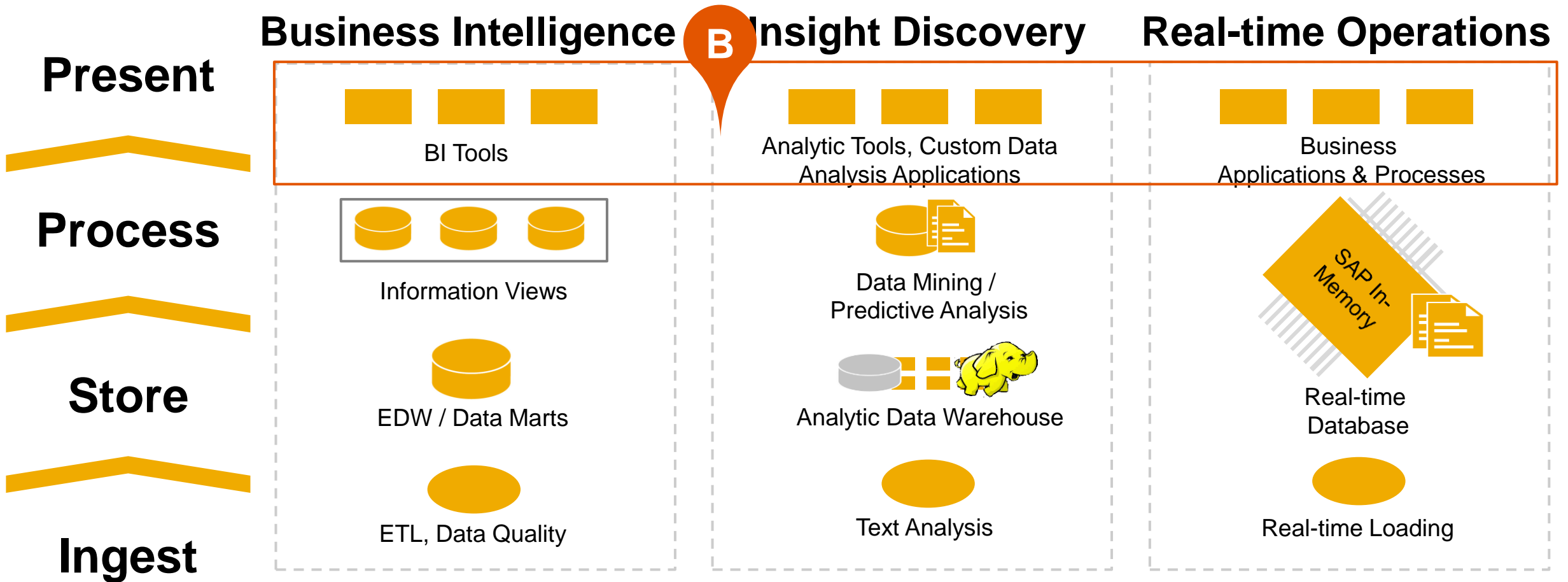
- ❑ Map text into pre-defined structures
- ❑ Cleanse, match, de-duplicate data

3 **Load** results quickly into EDW

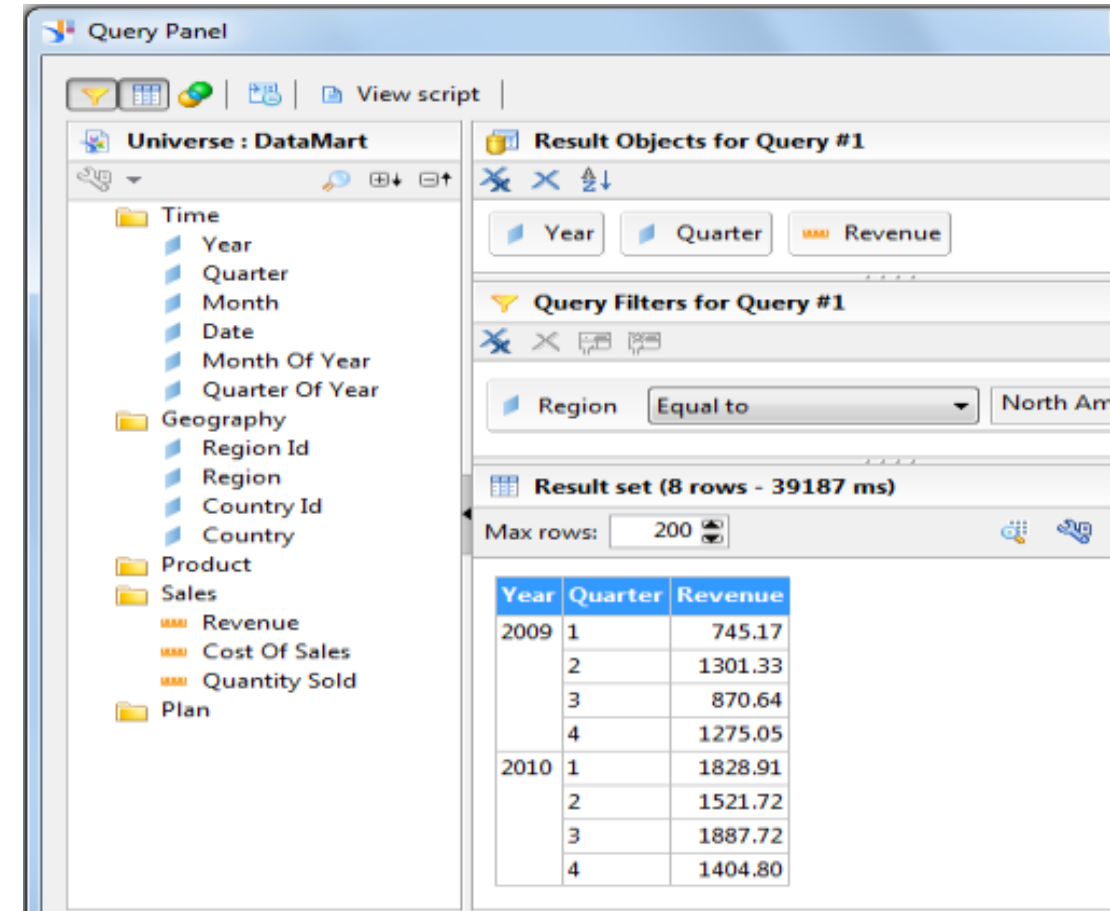
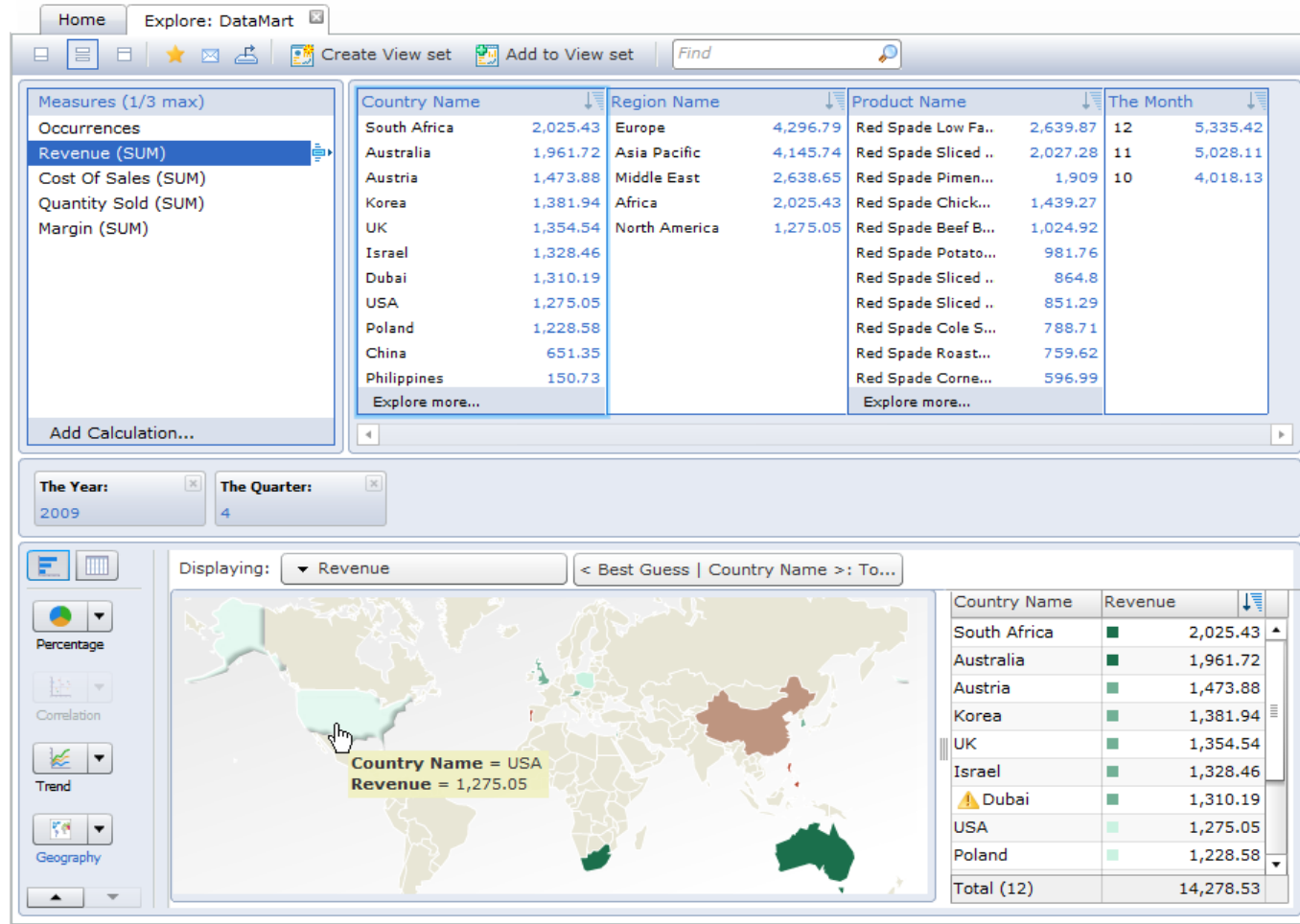
- ❑ Map text to structure



Hadoop Analytics



Business Analyst: Viewing data in Hadoop using GUI tools



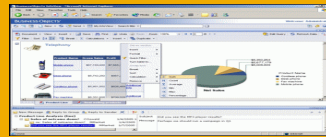
Automatically generates HiveQL statements that are executed on a Hadoop cluster

SAP BusinessObjects BI: Hadoop for Business Analysts



Common user experience for all front-end tools

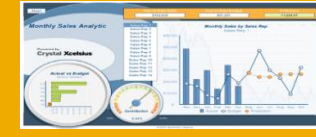
Empower all analysts,
enable all workflows



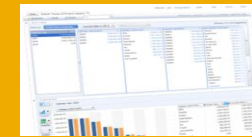
Web Intelligence



Crystal Reports



Dashboards



Explorer

Best access method for each specific data source

High performance,
feature rich, secure

Universe Access

Direct Access

All data sources

Extract, define, &
manipulate metadata



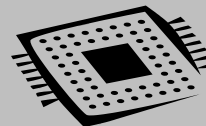
**HADOOP
HIVE**



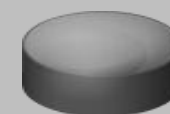
SAP BW



Sybase
databases



SAP HANA



3rd party
databases



Files

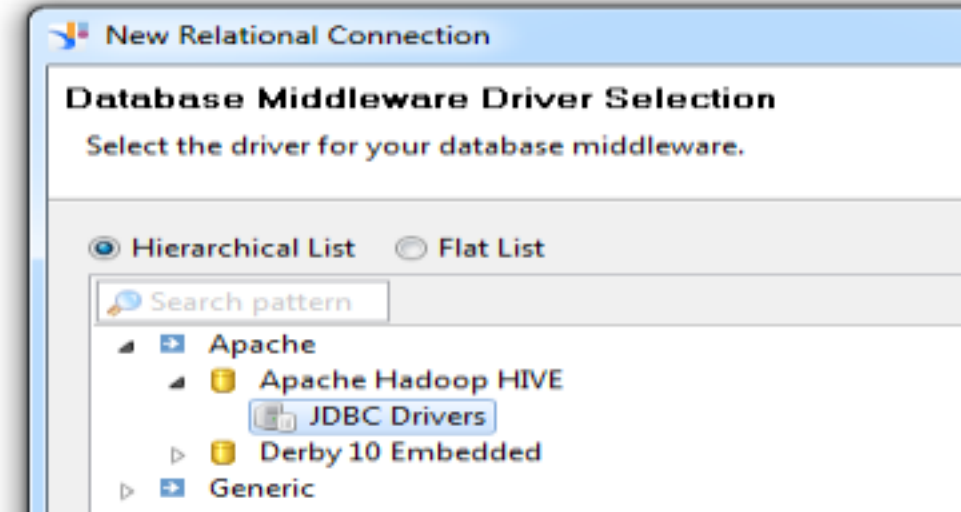
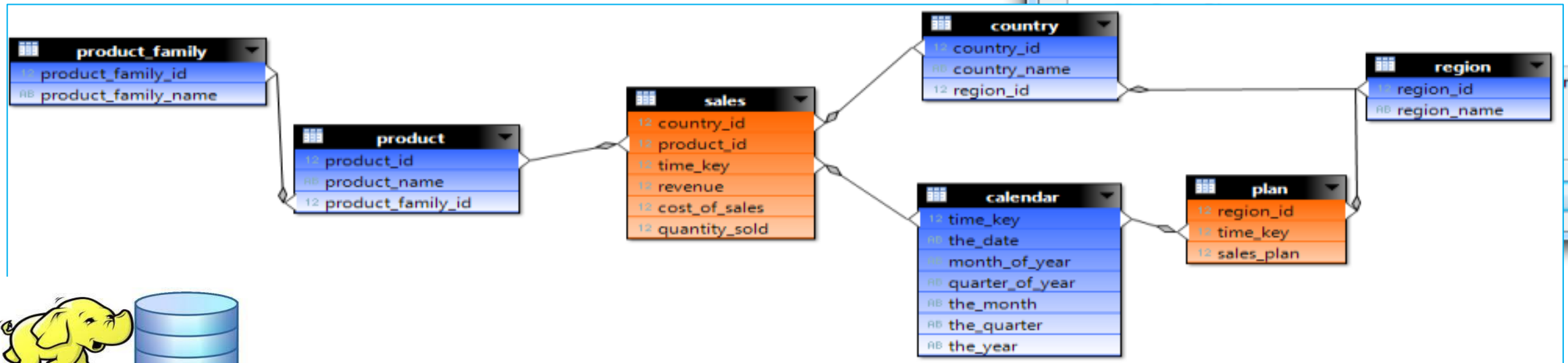


Web
Services

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



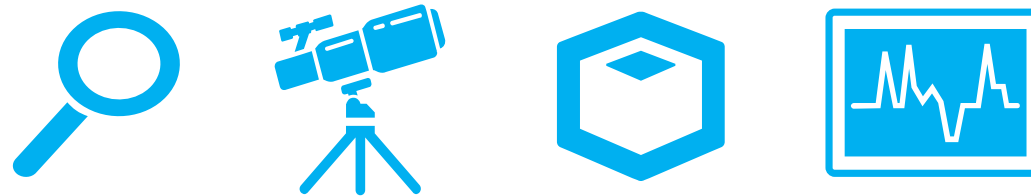
Data Scientist: Flexibility is of the essence



Chooses the variables that offer the most promise



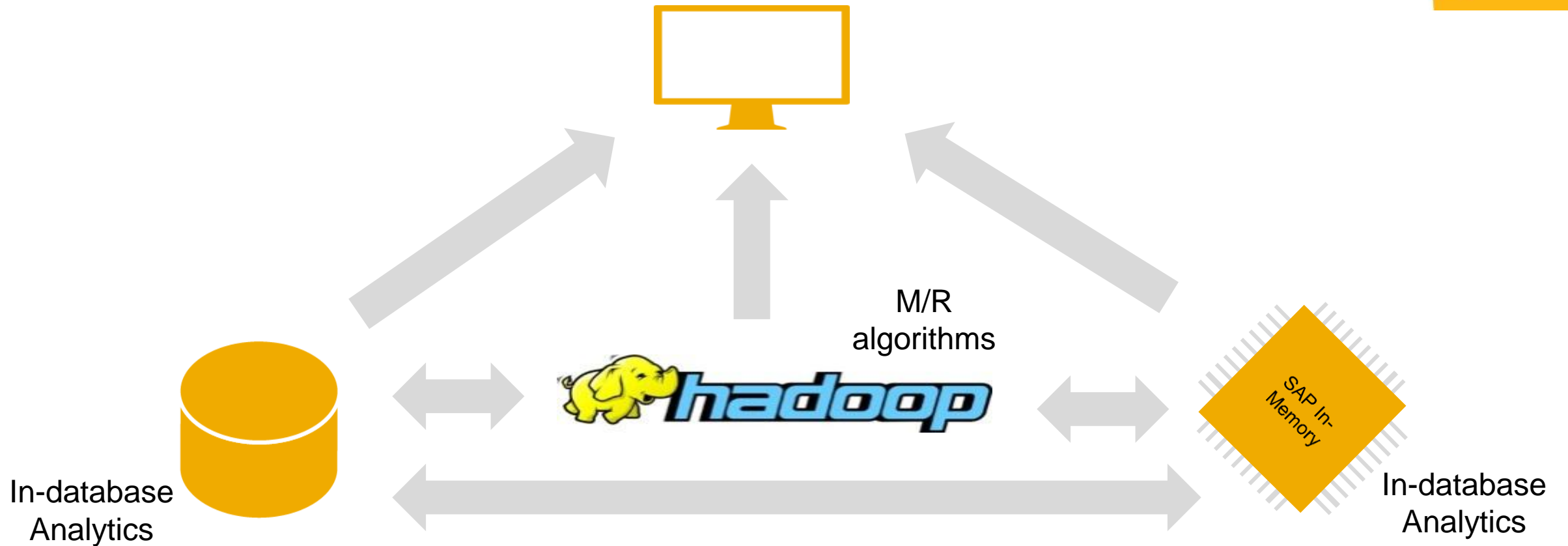
Chooses the best tool based on data mining technique



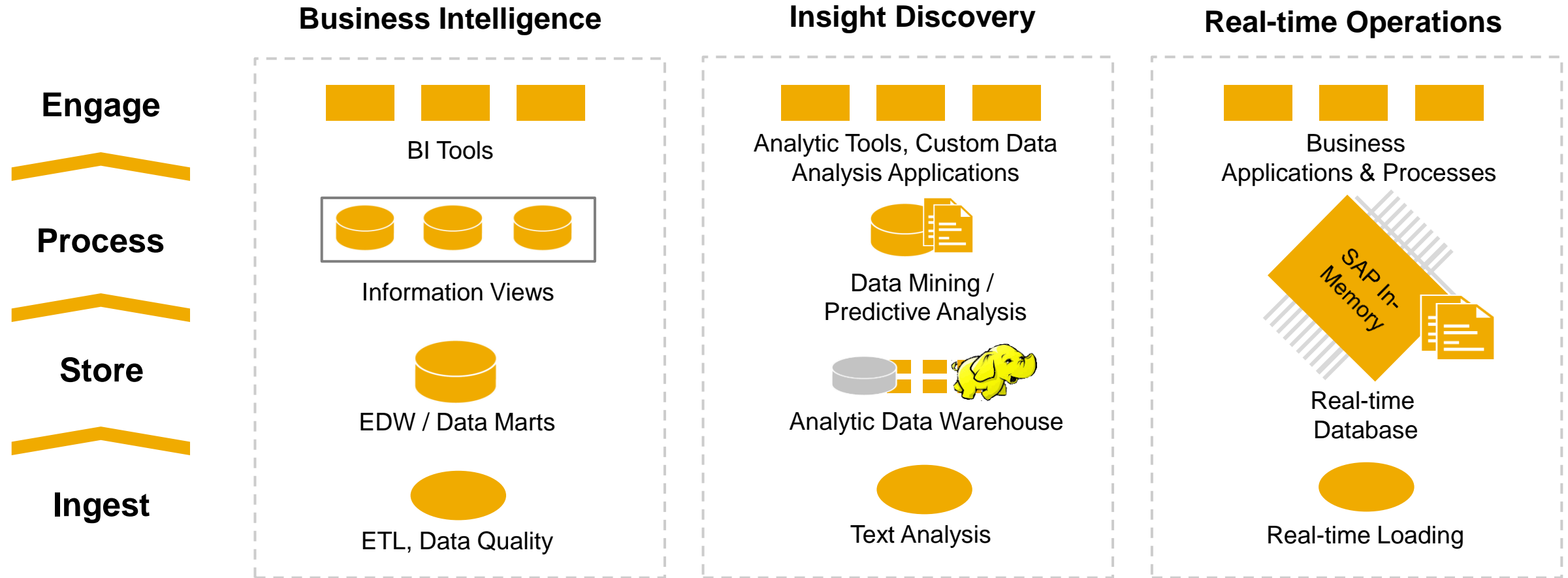
Chooses best analysis engine based on algorithm & data



Data Scientist: Open integration is key



Building an IT landscape for Big Data





Questions?
william.gardella@sap.com

© 2012 SAP AG. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

Microsoft, Windows, Excel, Outlook, PowerPoint, Silverlight, and Visual Studio are registered trademarks of Microsoft Corporation.

IBM, DB2, DB2 Universal Database, System i, System i5, System p, System p5, System x, System z, System z10, z10, z/VM, z/OS, OS/390, zEnterprise, PowerVM, Power Architecture, Power Systems, POWER7, POWER6+, POWER6, POWER, PowerHA, pureScale, PowerPC, BladeCenter, System Storage, Storwize, XIV, GPFS, HACMP, RETAIN, DB2 Connect, RACF, Redbooks, OS/2, AIX, Intelligent Miner, WebSphere, Tivoli, Informix, and Smarter Planet are trademarks or registered trademarks of IBM Corporation.

Linux is the registered trademark of Linus Torvalds in the United States and other countries.

Adobe, the Adobe logo, Acrobat, PostScript, and Reader are trademarks or registered trademarks of Adobe Systems Incorporated in the United States and other countries.

Oracle and Java are registered trademarks of Oracle and its affiliates.

UNIX, X/Open, OSF/1, and Motif are registered trademarks of the Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, and MultiWin are trademarks or registered trademarks of Citrix Systems Inc.

HTML, XML, XHTML, and W3C are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

Apple, App Store, iBooks, iPad, iPhone, iPhoto, iPod, iTunes, Multi-Touch, Objective-C, Retina, Safari, Siri, and Xcode are trademarks or registered trademarks of Apple Inc.

IOS is a registered trademark of Cisco Systems Inc.

RIM, BlackBerry, BBM, BlackBerry Curve, BlackBerry Bold, BlackBerry Pearl, BlackBerry Torch, BlackBerry Storm, BlackBerry Storm2, BlackBerry PlayBook, and BlackBerry App World are trademarks or registered trademarks of Research in Motion Limited.

Google App Engine, Google Apps, Google Checkout, Google Data API, Google Maps, Google Mobile Ads, Google Mobile Updater, Google Mobile, Google Store, Google Sync, Google Updater, Google Voice, Google Mail, Gmail, YouTube, Dalvik and Android are trademarks or registered trademarks of Google Inc.

INTERMEC is a registered trademark of Intermec Technologies Corporation.

Wi-Fi is a registered trademark of Wi-Fi Alliance.

Bluetooth is a registered trademark of Bluetooth SIG Inc.

Motorola is a registered trademark of Motorola Trademark Holdings LLC.

Computop is a registered trademark of Computop Wirtschaftsinformatik GmbH.

SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, SAP HANA, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Business Objects Software Ltd. Business Objects is an SAP company.

Sybase and Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere, and other Sybase products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Sybase Inc. Sybase is an SAP company.

Crossgate, m@gic EDDY, B2B 360° , and B2B 360° Services are registered trademarks of Crossgate AG in Germany and other countries. Crossgate is an SAP company.

All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

The information in this document is proprietary to SAP. No part of this document may be reproduced, copied, or transmitted in any form or for any purpose without the express prior written permission of SAP AG.

Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, ohne die ausdrückliche schriftliche Genehmigung durch SAP AG nicht gestattet. In dieser Publikation enthaltene Informationen können ohne vorherige Ankündigung geändert werden.

Die von SAP AG oder deren Vertriebsfirmen angebotenen Softwareprodukte können Softwarekomponenten auch anderer Softwarehersteller enthalten.

Microsoft, Windows, Excel, Outlook, und PowerPoint sind eingetragene Marken der Microsoft Corporation.

IBM, DB2, DB2 Universal Database, System i, System i5, System p, System p5, System x, System z, System z10, z10, z/VM, z/OS, OS/390, zEnterprise, PowerVM, Power Architecture, Power Systems, POWER7, POWER6+, POWER6, POWER, PowerHA, pureScale, PowerPC, BladeCenter, System Storage, Storwize, XIV, GPFS, HACMP, RETAIN, DB2 Connect, RACF, Redbooks, OS/2, AIX, Intelligent Miner, WebSphere, Tivoli, Informix und Smarter Planet sind Marken oder eingetragene Marken der IBM Corporation.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und anderen Ländern.

Adobe, das Adobe-Logo, Acrobat, PostScript und Reader sind Marken oder eingetragene Marken von Adobe Systems Incorporated in den USA und/oder anderen Ländern.

Oracle und Java sind eingetragene Marken von Oracle und/oder ihrer Tochtergesellschaften.

UNIX, X/Open, OSF/1 und Motif sind eingetragene Marken der Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame und MultiWin sind Marken oder eingetragene Marken von Citrix Systems, Inc.

HTML, XML, XHTML und W3C sind Marken oder eingetragene Marken des W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

Apple, App Store, iBooks, iPad, iPhone, iPhoto, iPod, iTunes, Multi-Touch, Objective-C, Retina, Safari, Siri und Xcode sind Marken oder eingetragene Marken der Apple Inc.

IOS ist eine eingetragene Marke von Cisco Systems Inc.

RIM, BlackBerry, BBM, BlackBerry Curve, BlackBerry Bold, BlackBerry Pearl, BlackBerry Torch, BlackBerry Storm, BlackBerry Storm2, BlackBerry PlayBook und BlackBerry App World sind Marken oder eingetragene Marken von Research in Motion Limited.

Google App Engine, Google Apps, Google Checkout, Google Data API, Google Maps, Google Mobile Ads, Google Mobile Updater, Google Mobile, Google Store, Google Sync, Google Updater, Google Voice, Google Mail, Gmail, YouTube, Dalvik und Android sind Marken oder eingetragene Marken von Google Inc.

INTERMEC ist eine eingetragene Marke der Intermec Technologies Corporation.

Wi-Fi ist eine eingetragene Marke der Wi-Fi Alliance.

Bluetooth ist eine eingetragene Marke von Bluetooth SIG Inc.

Motorola ist eine eingetragene Marke von Motorola Trademark Holdings, LLC.

Computop ist eine eingetragene Marke der Computop Wirtschaftsinformatik GmbH.

SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, SAP HANA und weitere im Text erwähnte SAP-Produkte und Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der SAP AG in Deutschland und anderen Ländern.

Business Objects und das Business-Objects-Logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius und andere im Text erwähnte Business-Objects-Produkte und Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der Business Objects Software Ltd. Business Objects ist ein Unternehmen der SAP AG.

Sybase und Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere und weitere im Text erwähnte Sybase-Produkte und -Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der Sybase Inc. Sybase ist ein Unternehmen der SAP AG.

Crossgate, m@gic EDDY, B2B 360° , B2B 360° Services sind eingetragene Marken der Crossgate AG in Deutschland und anderen Ländern. Crossgate ist ein Unternehmen der SAP AG.

Alle anderen Namen von Produkten und Dienstleistungen sind Marken der jeweiligen Firmen. Die Angaben im Text sind unverbindlich und dienen lediglich zu Informationszwecken. Produkte können länderspezifische Unterschiede aufweisen.

Die in dieser Publikation enthaltene Information ist Eigentum der SAP. Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, nur mit ausdrücklicher schriftlicher Genehmigung durch SAP AG gestattet.