Deloitte.

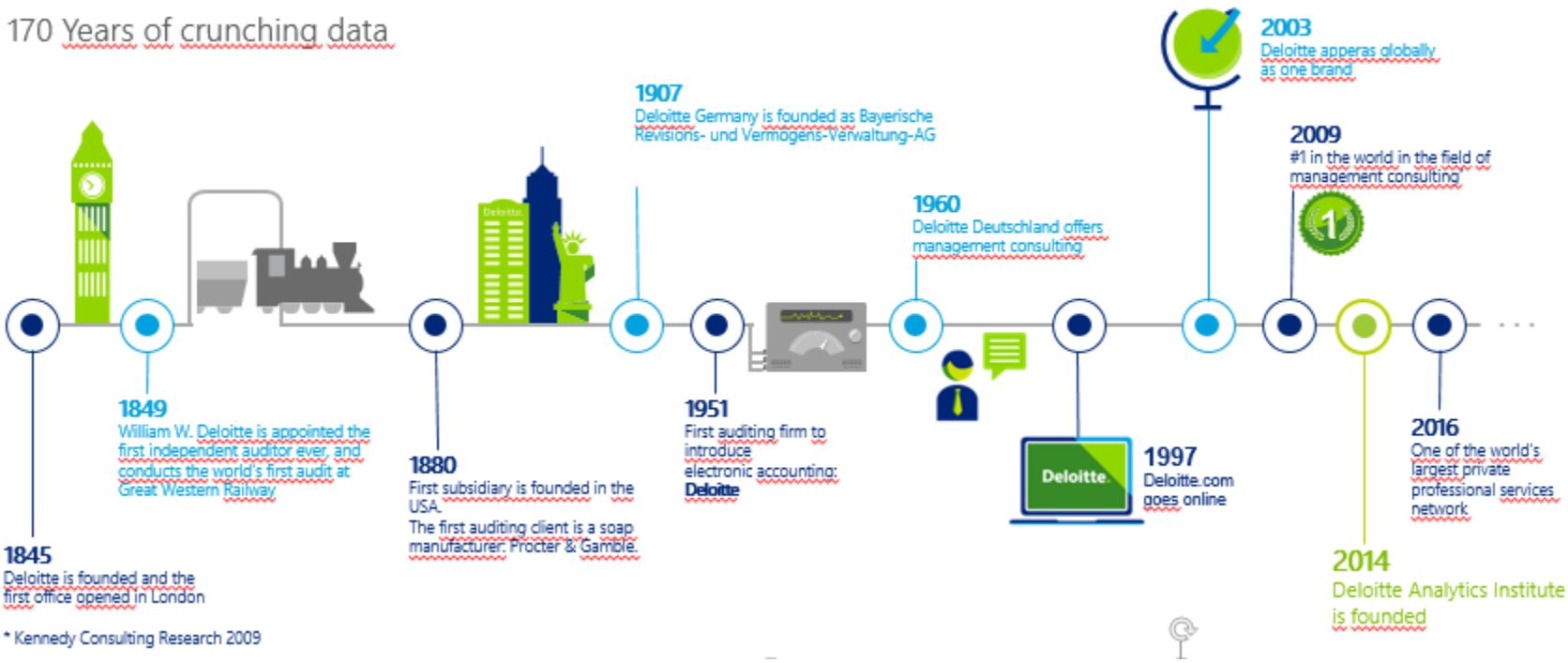
Big Data Architectures

Lessons Learned from Industrializing Big Data

Kenan Mujkic, PhD
23 June 2016

analytics institute

Deloitte



Making an impact that matters – for clients, for our people, and for society. We serve clients distinctively, bringing innovative insights, solving complex challenges and unlocking sustainable growth. We inspire our 220,000 talented professionals to deliver outstanding value.

Deloitte provides audit, tax, financial advisory and consulting services to public and private clients spanning multiple industries; legal advisory services in Germany are provided by Deloitte Legal. With a globally connected network of member firms in more than 150 countries, Deloitte brings world-class capabilities and high-quality service to clients.

Contents

Scope of Big Data Projects

Scope of Big Data Projects

What do clients expect from Big Data?

1. **Business Insights**
   Leverage business value from advanced analytics capabilities, e.g. enhanced online marketing, branding and (re)targeting using customer segmentation.

2. **Business Models**
   Novel product lines originating from the digital transformation of the economy, e.g. insurance rates based on the scoring originating from telematics data.

3. **Reduced Storage Cost**
   Reduction of cost from storage of business relevant data in HDFS, and enhanced computing performance for data pre-processing using parallel computing.

4. **Industry 4.0, Internet of Things**
   Enhanced productivity and reduced defects in production lines, e.g. for consumer products produced based on data driven decision making, predictive maintenance.

5. **Enhanced Service and Product Quality**
   Enhanced service and product quality due to predictive maintenance solutions in automotive, railway, aerospace and other industries.

Scope of Big Data Projects

What are typical Big Data architectures?

Kappa Architecture

Architecture: Streaming layer (Spark, Storm) for real-time analytics and decisioning.

Purpose: Real-time analytics, scoring results stored in serving layer for visualisation and analytical purposes.

Workload: Large workloads, relatively simple analytics use cases.

Lambda Architecture

Architecture: Real-time layer (Spark, Storm), batch layer (MapReduce/Hive) and serving layer (HBase, MapR DB, Cassandra, etc.).

Purpose: Combination of historical and transactional data in serving layer.

Workload: Large workloads, complex real-time & batch processing use cases.

Batch Architecture

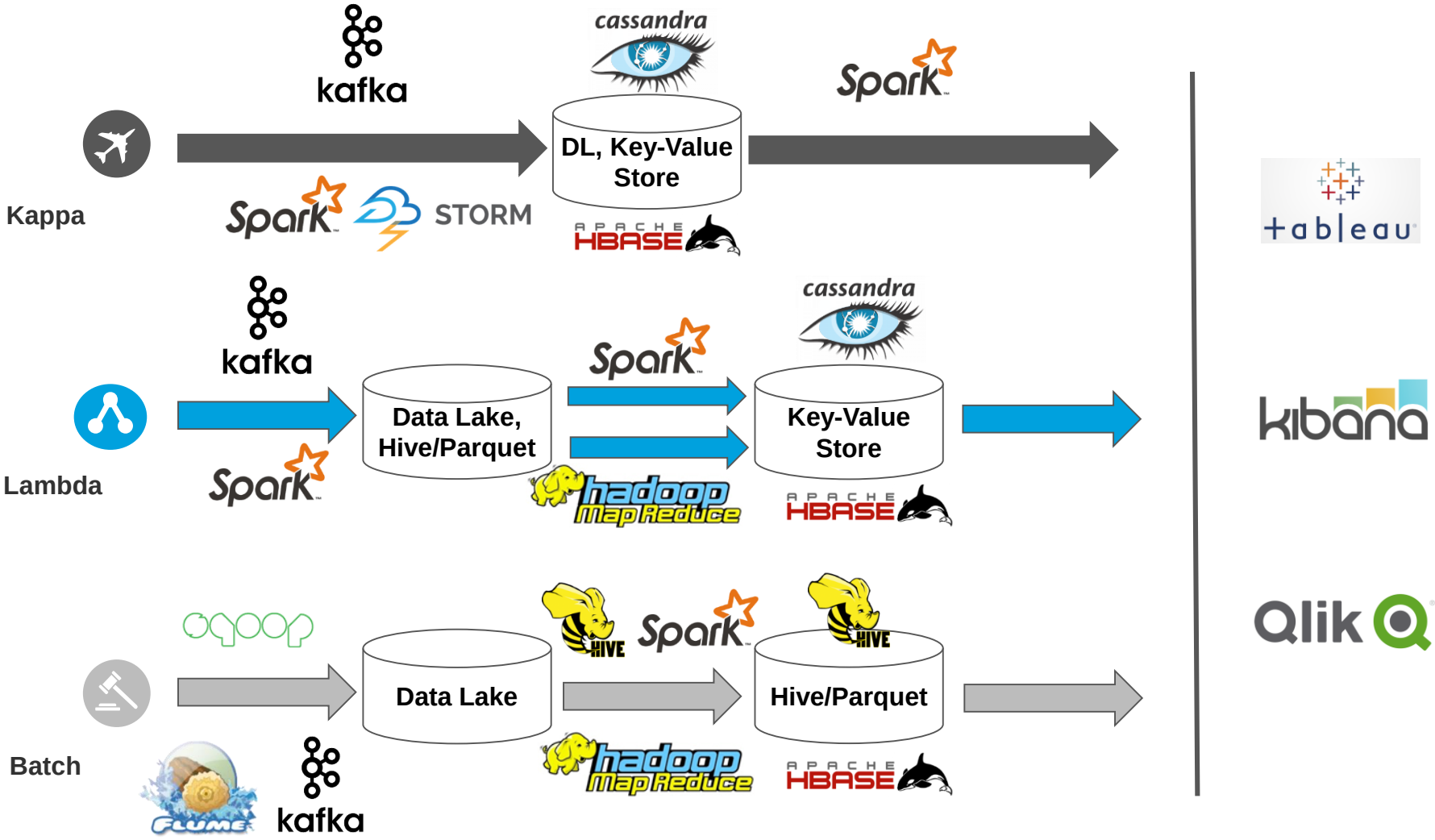
Architecture: Batch layer (Hive/Parquet) for storage of large data sets.

Purpose: Classical storage of historical data for analytical purposes using batch processing.

Workload: Large workloads, complex analytics use cases.

Scope of Big Data Projects

What are typical Big Data architectures?



Deloitte - Selected Projects

Big Data Project Examples

Selection of some of our current Big Data Platform projects at large companies in Germany

Financial Service Provider – European BI Platform

- **Scope:** European platform for regulatory reporting
- **Purpose:** Perform regulatory reporting according to European standards, leverage and combine current as well as historical data from a variety of sources, develop new use cases

Finance

Insurance Company – Global Telematics Platform

- **Scope:** Scalable platform for real-time trip processing and scoring
- **Purpose:** Provide customers with individual insurance rates, based on their driving behavior, real-time ingestion of trip data, crash data and harsh events, mobile app integration

Insurance

Automotive Manufacturer – Global Big Data Platform

- **Scope:** Ingestion, storage and analysis of data from vehicle and production devices
- **Purpose:** Identify location and speed to improve traffic information in cars, collect & analyze sensor information from production robots and workstations to decrease failure rates

Automotive

Financial Service Provider

European BI & Reporting Platform

Issue

- New European reporting and regulation structure
- Lack of standardization of risk data across sources, single stop for all risk related data does not exist
- Governance processes around the risk data are insufficient

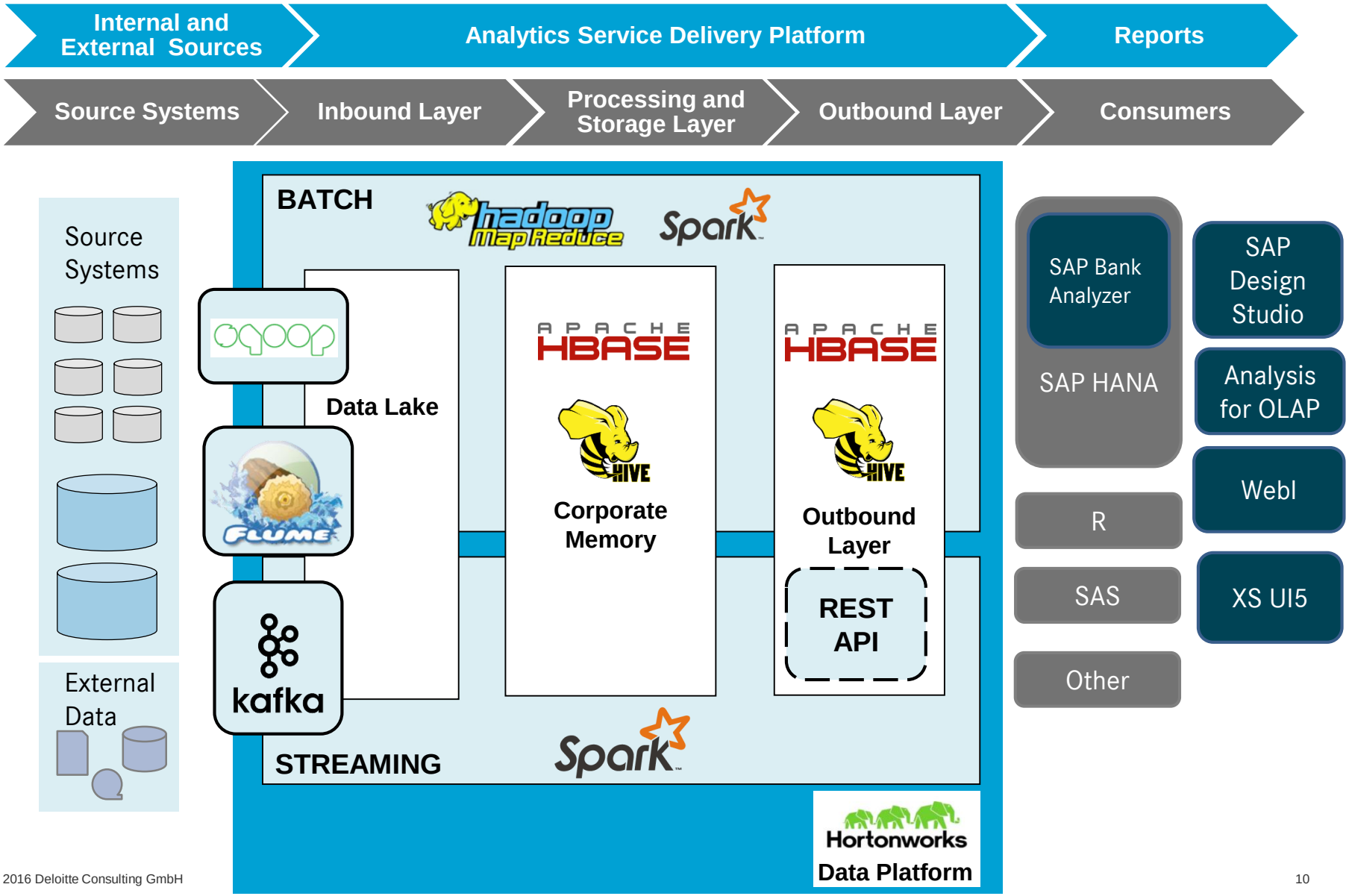
Solution

- Migration of SAP BW to HANA
- Leverage the storage capabilities and performance of Hadoop and the analytical power of SAP Bank Analyzer on Hana
- Migration of ETL from SAP BW Tool to Spark/Hadoop
- Standardized source systems for all subsidiaries

Impact

- Standardized regulatory reporting and risk data from multiple sources
- Consolidated business data stored in a unified corporate memory
- Enable reporting and visualization of risk data across the functions Investment Risk, Counterparty Risk and Operational Risk

Financial Service Provider – European BI Platform Architecture



Insurance Company

Global Telematics Platform

Issue

- Changes on agreement level and tariffs are important requirements for the near future.
- No detailed and standardized information about the customer's master data and driving behavior are available.
- Manual consolidation processes and management reports with limited ad-hoc reporting.

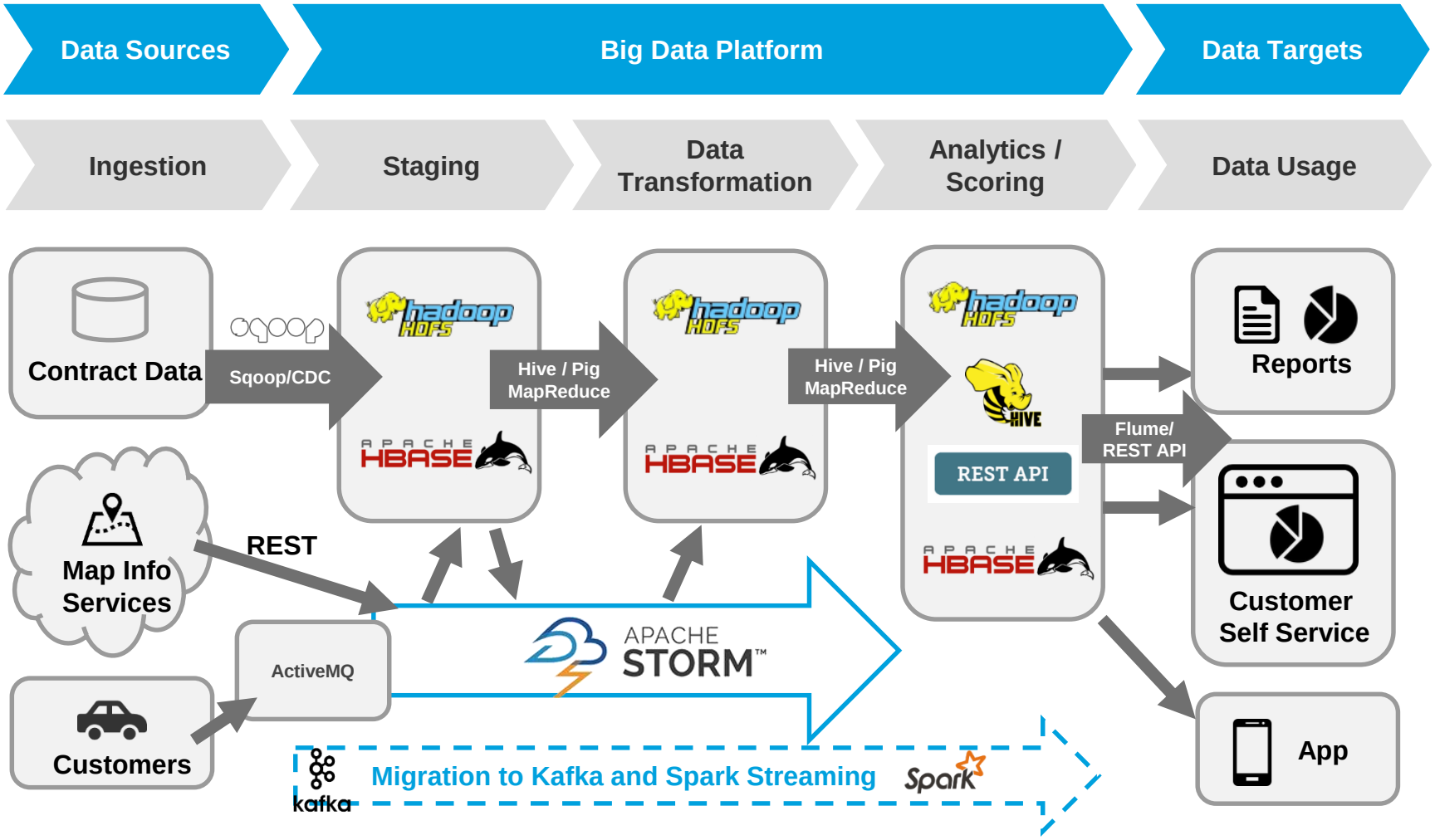
Solution

- Leverage the real-time processing capabilities of Apache Storm.
- Take advantage of the storage capabilities and high performance of Hadoop and the analytical and visualization capabilities of R and Tableau.
- Scope: Real-time analytics and scalability for trip processing and batch processing for scoring to calculate individual insurance rates, based on the driving behavior of the customer.

Impact

- New offerings for insurance leads and customers.
- Standardized trip data processing to assess the driving behavior of customers, relating to consolidated and standardized master data.
- Enable enhanced reporting and visualization of customer's driving behavior, and one consolidated view of risk data to the business user.

Insurance Company – Telematics Platform Architecture



Automotive Manufacturer

Global Big Data Platform

Issue

- Customer satisfaction and sales offers are a future improvement goal.
- Vehicle data is essential to analyze and predict near real-time, lack of standardization of contract, finance and vehicle data across sources.
- Connected car data needs faster central analytics for more valuable services and proactive offers and announcements.

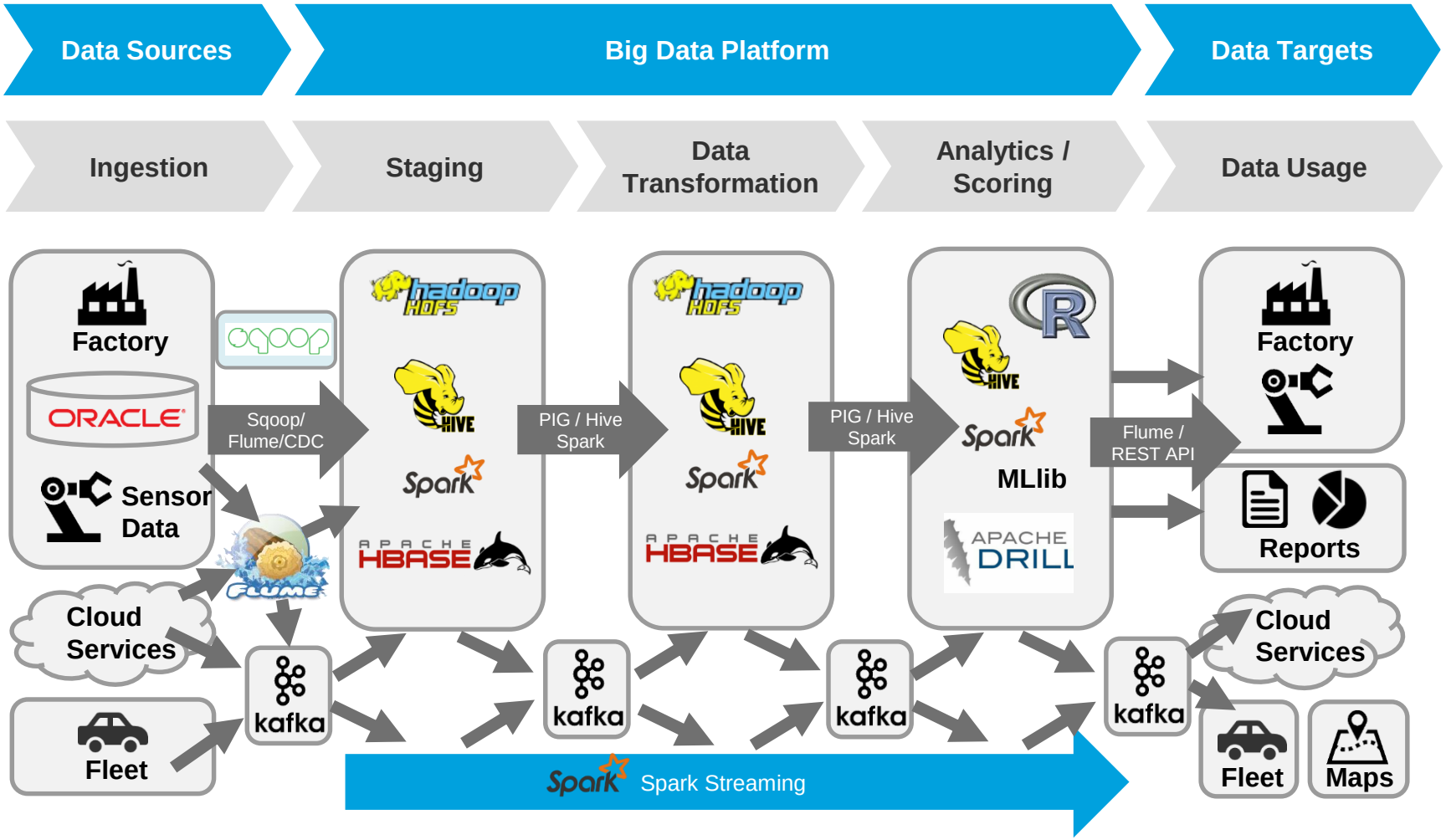
Solution

- Leverage the analytics capabilities of Spark, and the storage and processing capabilities of Hadoop.
- Take advantage from a structured data warehousing approach on HDFS, with visualization, analytics and reporting based on QlikView, Tableau and R.
- Enhance the capabilities of Hadoop by enabling streaming, real-time processing, time series and predictive analytics.

Impact

- Standardized data from multiple sources.
- Enable real-time streaming, replication and analytics of vehicle data using Spark, Kafka and Hadoop.
- Enhanced reporting & visualization of customer, vehicle and trip related data.
- A shared data delivery ecosystem for every involved department and production plant.

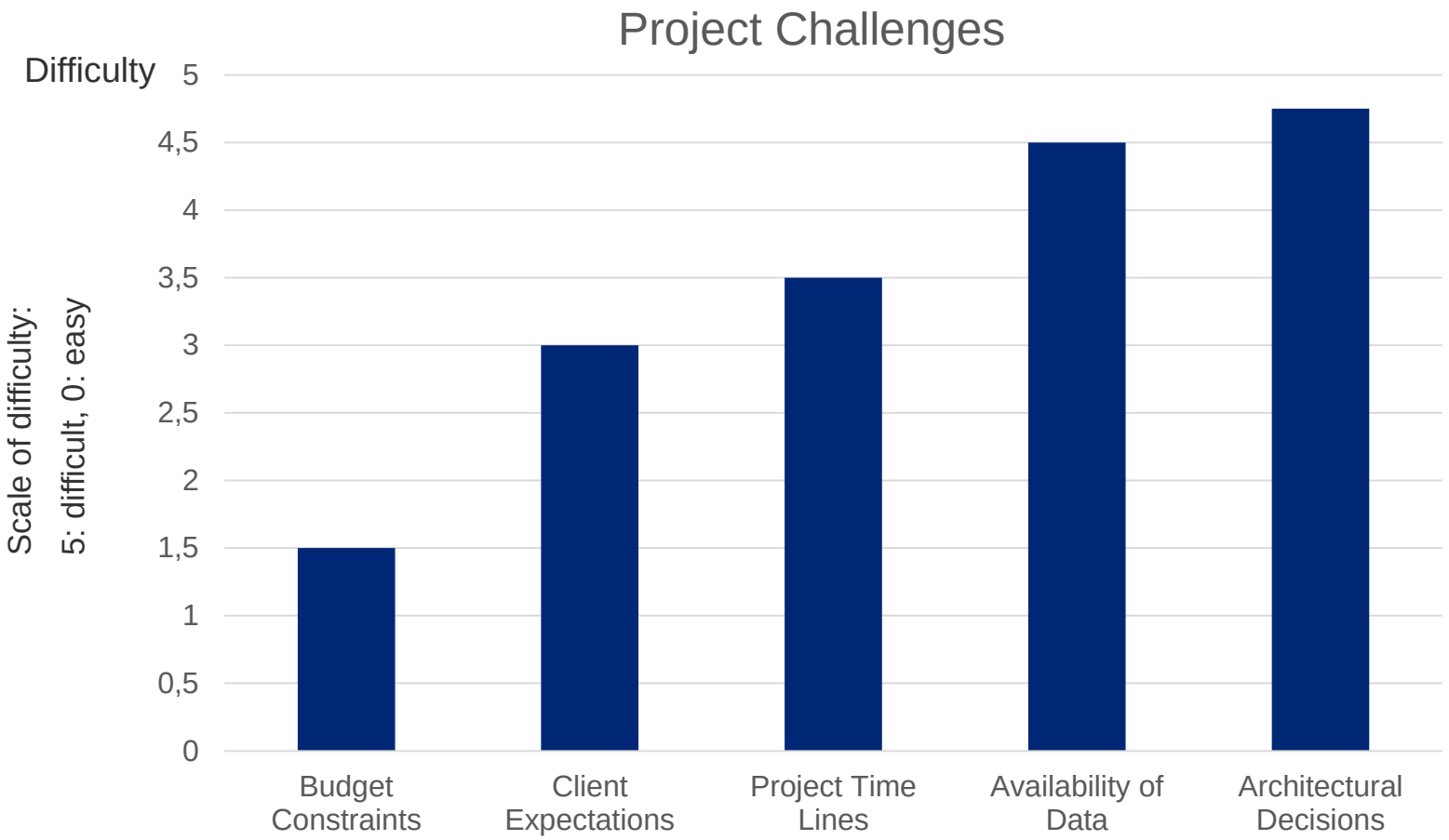
Automotive Manufacturer – Big Data Platform Architecture



Industrializing Big Data – Lessons Learned

Industrializing Big Data - Lessons Learned

What are the main challenges in Big Data projects?



Industrializing Big Data - Lessons Learned

What are the main technical challenges in Big Data projects?

Configuration of Hadoop Cluster

- YARN Container Size
- Oozie / Hive-Impersonation

Kerberos Authentication

Multi-Tenancy, Authorisation and Security

- Authorisation Policies and Groups

Hive-HBase Dichotomy

- Hadoop is not a RDBMS
- Small Files Problem / Compaction

HBase

- Hashing and Distribution
- Versioning

Spark

- Distributed Cluster Mode
- Security

Development / Deployment of Applications

- Jira / Confluence / VMs
- Jenkins, Maven, SVN, Git
- Eclipse, IntelliJ IDEA

Automated Testing of Applications

- Junit Tests
- System & Integration Tests

Automation of Applications

- Scheduling & orchestration: Oozie, Spark, Cron, etc.

Industrializing Big Data - Lessons Learned

What are the main pitfalls in Big Data projects?

Business Requirements

- Missing or frequently changing business requirements lead to flawed architectures.

Technical Skills of Project Management

- Project managers often underestimate the effort to configure Hadoop clusters extending project time lines.

Data Quality and Consistency

- Data samples that are not representative or entirely missing data sets lead to extended project time lines.

Scalability and Small Files

- Performance and scalability tests with large data sets on Hadoop Cluster, small files compaction

Management of Client Expectations

- Many clients expect that Big Data Platforms are more efficient, less expensive and easier to handle than classical Enterprise Data Warehouses (EDW).

- In reality, Hadoop clusters are usually more difficult to configure and administrate, a.o. due to the lacking skill set in the IT departments.

- For this reason it is crucial to manage client expectations at the beginning of a Big Data project to achieve the project goals and meet important deadlines.

Thank you very much for your attention!