Future-proof architecture for a modern data platform

SOLITA

Table of contents

Foreword

Business needs are evolving rapidly. Digitalization, Internet of Things (IoT) and social media are creating a vast variety of information that companies would like to take advantage of and build business models to fully utilize the data.

Many traditional data warehouses are challenged with the requirements around modernization, as big data with real-time analytics demands a new way of handling data. With this in mind, the need for a flexible, reliable and scalable data platform is pressing, and it is time for enterprises to adopt modern cloud platforms.

This whitepaper showcases data platform architecture in the cloud and describes the needed elements to make up the architecture. The details provide an understanding on how these architecture components enable enterprise class platforms with massive performance and scale that is not achievable by legacy business intelligence and data warehouse platforms.

Data-driven business requires new data architecture

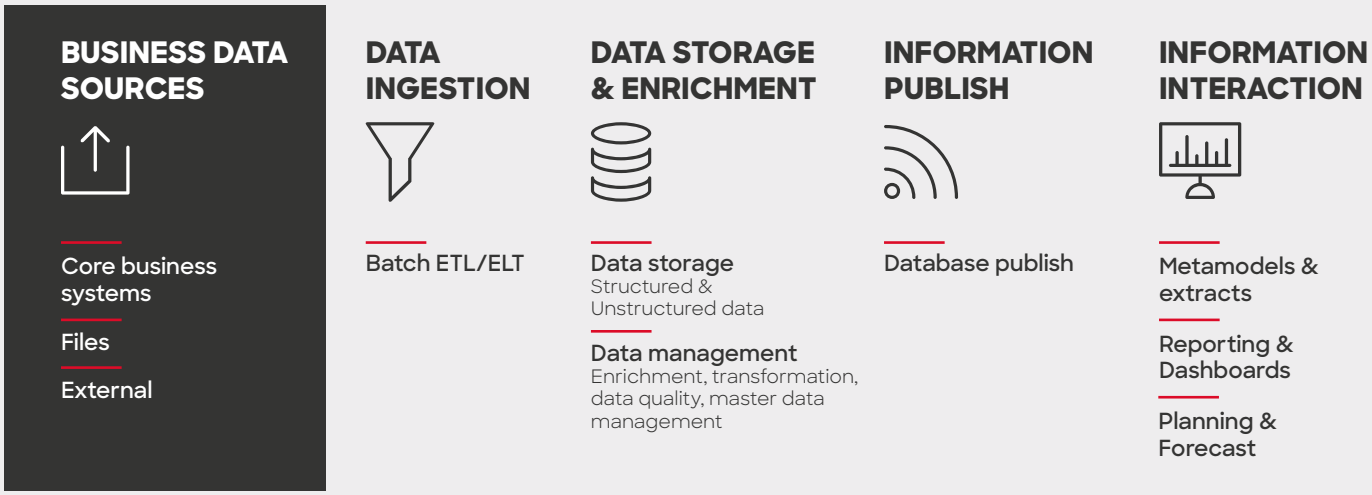
Data has become central to how businesses make decisions. Many companies have always used data, but as we are able to collect more data at higher speed and more cost efficiently, this has a huge impact on how data management needs to adapt.

Doug Laney coined the term big data with three V's, adding a fourth one later: volume, variety, velocity and veracity push traditional data management to the edge. In order to adapt to these changes, new agile ways of managing data are needed. Data needs to be processed at the right time, stored for analytical usage and made available for data consumers all while putting governance on it to make sure the data is of high quality. Infrastructure of the processing systems need to scale to the needs of the data as well as the business.

If these demands were not enough, in many cases we are dealing with sensitive data, and have to make sure we are aligned with rules and regulations as well as maintain control of our important data assets. With the new requirements, we need new ways to take on these challenges.

Legacy business intelligence and data warehousing system architectures comprise of taking data from various structured sources (both internal and external), batch ingestion of source data, data storage and management including transformation and enrichment within data warehouse (RDBMS) and creation of reports and dashboards as per business requirements with the help of the publish layer in the warehouse.

Below is a simple representation of a legacy Business Intelligence and Data Warehousing architecture:



These systems are limited in their capacity and scalable up to a certain extent, mainly being on-premise and created for structured data use cases. It is challenging for these systems to be extended to handle modern use cases like streaming and IoT data, making them auto-scalable as per demands, and incorporating use cases like Machine Learning and Artificial Intelligence.

As enterprises are looking for architecture that supports the use cases of modern business, it is important to ensure system architecture which could handle it and is flexible and scalable as per the needs, and hence the evolution of data platform based on public cloud platform is important to understand.

What is a data platform?

Data platform is a technical solution to enable value creating interactions between different data producers and data consumers. The platform provides infrastructure for these interactions, management of data, ways to improve data quality, and sets governance conditions for the data asset.

Data platforms should enable value creation in a simplified manner for both producers and consumers of data. Enterprises usually deal with a lot of sources and destinations, and data platforms should not be a bottleneck. It is important to ensure a simple process for producers who want their data incorporated into the system and for consumers who want to utilize the data. To understand this, we could draw an analogy to the simplicity of listing and booking through Airbnb. Anyone can list their accommodation on the platform without much preparation, and anyone can book an accommodation without much planning. This seamless interaction is what modern platforms should aim to achieve.

Reference: Platform Revolution by Geoffrey G. Parker, Marshall W. Van Alstyne and Sangeet Paul Choudhary

Data platform capabilities and architectural components

The platform should be built alongside the use cases in an agile manner since new use cases pop up all the time. Thinking about system architecture needs to be holistic. Building data platforms should not be an IT exercise. Business teams need to be onboarded early on in the process when selecting and working with relevant business use cases in order to create value for the organization. See below for a few customer use cases we are actively working with:

EXAMPLE 1

One of our customers has immensely increased their data capabilities by combining data from existing sources like digital equipment, mobile applications and sales database to create a comprehensive customer profile, used extensively around the company from product design to highly targeted marketing and sales predictions. This use case touches on components for sourcing batch data, data storage, data analytics and publish with feedback to various sources and departments. New use cases for data and exploration of additional components to data platform is happening continuously through agile testing periods.

EXAMPLE 2

Our customer started with the aim to develop real-time capabilities instead of conventional financial reporting so that the data can be shared, combined and requested easily. Instead of business intelligence, they can treat data as an asset and utilize it. This acts as a solid foundation that allows the creation of new use cases, since data from diverse sources is already made available in the platform and is available for potential new business use cases.

EXAMPLE

3

For companies of the future, one cornerstone of success is their ability to use data in a versatile manner to create new operating models, services and business. Our customer was looking to collect and combine data about their customers in a more flexible manner for segmentation and creation of new services. A better, cost-effective way of using the data required a modern data warehouse that could combine data from different sources into one ecosystem and allow the development of new, data-driven services.

The above use cases utilise the following architectural components in their data platform journeys. The main components of data platform architecture are:

The interactions from data producers

Data storage & enrichment

Business data sources & channels

Information publish

Data ingestion

Information interaction with data consumers

The cornerstones that support this architecture for it to be built and maintained effectively and in an agile and scalable manner are:

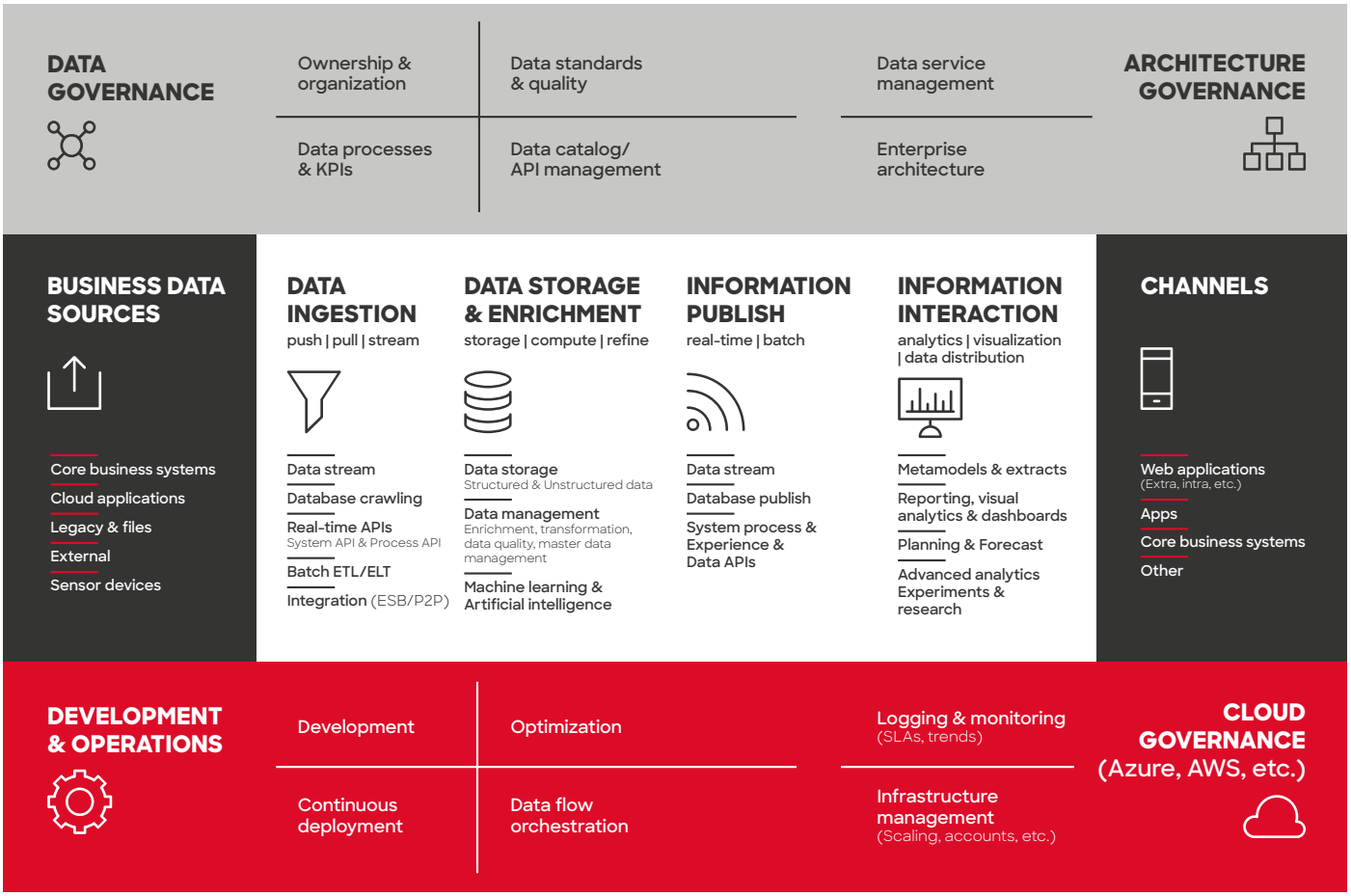
Data governance

Architecture governance

Development & operations

Cloud governance

> **One should use the best tools for the job, and this architecture presents different parts that might be useful. A lot can be done without them all being in place.**



Data ingestion

Data ingestion is about moving data from where it originated into a system where it can be stored and analysed. Data can be streamed in real time or ingested in batches.

Stream data is data that is continuously generated by different sources, which typically send in the data records simultaneously, and in small batch sizes.

Crawlers are a set of techniques aiming to automate the collection of content from WWW, databases and other data sources.

API (Application Programming Interface) is a software intermediary that allows two applications to talk to each other. More and more data (e.g. web-based SaaS services) are ingested to data platforms via application APIs.

ETL (Extract, Transform and Load): Three functions of a data integration process that are combined to pull data from a source and place it in a destination database. Transformation takes place on an intermediate server before it is loaded into the target.

ELT is a variation of ETL; it leverages the target system to do the transformation. The data is copied to the target and then transformed in place. This is becoming a popular transformation method with Cloud Platform evolution.

An **ESB** (enterprise service bus) is primarily to provide the connections between communicating applications, acting much like a router to control the data. It is commonly used in enterprise application integration (EAI) principles.

Data storage & enrichment

Data storage and enrichment comprise of all the disciplines related to managing and enrichment of data as a valuable resource.

Modern platforms should be capable of storing and processing both structured and unstructured data. A set of tools and techniques are needed to store information in the correct format. **Structured data storage** is information that has a predefined data model or is organized in a predefined manner. **Unstructured data storage** does not have a predefined data model and is not organized in a predefined manner, an example being images and videos, etc.

Data compute refers to collections of techniques and platforms used to process data to various forms. Generally, this means cloud computing: ubiquitous access to shared pools of configurable system resources and higher-level services that can be rapidly provisioned with minimal management effort, often over the internet.

Machine learning (ML) and **Artificial intelligence (AI)** is used to learn about data in an automated way. They are used for software applications to become more accurate in predicting outcomes without being explicitly programmed, learning through sample and historical data.

Master data management (MDM) is about improving and ensuring the quality of data and making it available for different users, systems and applications at a correct time.

Information publish

Information publish layer is the uniform information model which contains business logic, and it is used by applications and services. Data could be in virtualized or materialized format.

Data virtualization is an approach that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted at source, or where it is physically located.

DBMS publish models refer to a collection of modeling techniques by which information can be published to use. Typically support end-user queries in a data warehouse and hide underlying data complexity. The most well-known publish model is dimensional modeling (star schema).

System APIs are interfaces which hide core systems complexity from the user while exposing data and providing downstream insulation from any interface changes or rationalization of those systems. **Process APIs** encapsulate the underlying business processes that interact with source and target systems or channels via a set of system APIs.

Information interaction

Consuming and interacting with the information in different ways; both internal and external to the enterprise.

Metamodels and extracts: Published Data is typically served for self-service users (analysts) and data professionals via software product specific metadata models or sometimes data extracts (e.g. in Tableau).

Visual analytics & dashboards & reporting: Data visualization and BI applications are still the most common type of using information provided by data platforms. Ready-made reports are consumed by end users that are not skilled enough to create their own reports and visualizations.

Planning & forecast: Systems and applications for creating operational and financial plans and forecasts (performance management). Typically, manual input is done by users and combined with actual figures to enable comparison of actuals vs targets.

Advanced analytics experiments & research: Data science and machine learning-based analysis research applications. The data scientist users need to have flexible access to data so they can do their research and ML training work in a custom way with different tools (R, Python, etc.). Some of the analytical models created in the research phase are then deployed to production.

Data governance

Data governance is a collection of practices and processes which help to ensure the formal management of data assets within an organization. How systematically data governance should be implemented in practice depends on the size, complexity and model of the organization and its maturity level. Having data governance practises in place gives the organization tools to monitor and improve data quality. Quality data creates trust which can be seen as a cornerstone of data-driven culture.

Ownership & organization: Ownership implies authority and responsibility. Owner of the data will be responsible for the timeliness and accuracy of the data, and can be contacted when questions arise about the data or its meaning and quality. From a data platform perspective it is important to know who is responsible for the source data and source data interfaces.

Data processes & KPIs: To implement the data governance in a bigger organization, clear processes are needed. One example of this is the data retention process which defines the policies of persistent data and records management for meeting legal and business data archival requirements.

Data standards & quality: Data standards are the rules by which data are described and recorded. In order to share, exchange and understand data, we must standardize the format as well as the meaning. Data quality refers to the condition of a set of values of qualitative or quantitative variables and needs to be actively monitored. Roles like Data Steward can support this.

Authentication & access control: All users need to be authenticated to be able to manage access to data on an individual and role basis. Typically, data platforms have many ways to implement this.

Usage logging: It is a good practice to gather data about what data is used and how. This usage metadata can be used to optimize performance, user experience and costs of the data platform.

Metadata management: End-to-end process and governance framework for creating, controlling, enhancing, attributing, defining and managing a metadata schema, model or other structured aggregation system, either independently or within a repository and the associated supporting processes.

Data catalog & API management: A data catalog is a metadata management tool designed to help organizations find and manage large amounts of data – including tables, files and databases – stored in their systems. Data catalogs centralize metadata in one location, provide a full view of each piece of data across databases, and contain information about the data's location, profile, statistics, summaries and comments. API management is the process of creating and publishing web APIs, enforcing their usage policies, controlling access, nurturing the subscriber community, collecting and analyzing usage statistics, and reporting on performance.

Architecture governance

Architecture of a data platform needs constant central governance as it serves many business purposes and projects, and it is not typically of main interest area for individual projects.

Data service management: It can be useful to think of the data platform as a combination of data services that are continuously evolving and have lots of dependencies. Data platform implements different use cases and has different workloads, which have varying non-functional requirements (security, performance etc.). So, it is important to design and implement a modular architecture to enable flexibility and resilience towards changing requirements and technologies.

Enterprise architecture (EA): A systematic approach for analyzing, designing and planning of organization activities to achieve its strategy. EA applies architecture principles and practices to guide organizations through the business, information, process, and technology changes necessary to execute their strategies.

Development operations

Development & operations contain a set of practices and ways of working to improve the quality and reduce the cycle time of data solution implementation, deployment and operations. Automation is a core target in this work, but it is also about emphasizing collaboration and integration between data scientists, data professionals and data engineers. This is where the emphasis is on DevOps and DataOps.

Development: Development practices need to be steered towards standardization and automation as much as possible to avoid person dependencies. Data testing is becoming an increasingly important piece of development.

Continuous deployment: Continuous integration and deployment pipeline/ environment should be established in all bigger data platform projects.

Optimization: It is important to continuously optimize the way of working in development and operations to improve both time-to-value and cost efficiency. Technologies are also developing all the time, so platforms need to be updated constantly.

Data flow orchestration: It would be ideal to have one centralized data flow orchestration tool/system for the whole data platform.

Cloud governance

Along with the development of the platform and its monitoring, it is crucial to govern the cloud environment that is hosting the data platform. There are several tools available which work both with and across cloud platforms and on-premise infrastructure.

Infrastructure management: With public cloud platforms, it is highly recommended to automate the infrastructure/service provisioning and related practices. If there are many people working within this area, the system can quickly become a mess, if not handled systematically.

Logging and monitoring: Logging and monitoring all technical activities in the environments is critical to be able to troubleshoot and optimize the data platform. There are typically several different logging components in the platform.

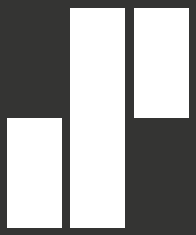
Summary

A modern data platform helps organizations avoid potential problems of data management by providing solutions to key challenges. With a more flexible structure and compatibility with leading-edge technology, the modern data platform is equipped to handle today's constantly changing data requirements. Its simplified architecture clears the path from data capture to actionable insight, thereby reducing decision lag. Crucially, it delivers these benefits without compromising on governance requirements.

In this document we presented the key components that make up the data platform. Components are chosen to support the organisations strategy in order to fulfill the business objectives. The platform should be driven by business cases, providing value and therefore return of investment, from the very beginning. This also engages business to the data platform, keeping the interest of stakeholders. The more mature and larger data platforms can compose all the components presented as they are tailored to the need. Regardless of your maturity with data management, this document can be used as a principle guide regarding the design of your data platform journey.

Annexure

AI	Artificial Intelligence
API	Application Programming Interface
BI	Business Intelligence
DW	Data Warehouse
EA	Enterprise Architecture
ELT	Extract Load Transform
ETL	Extract Transform Load
ESB	Enterprise Service Bus
IAAS	Infrastructure as a Service
IOT	Internet of Things
IT	Information Technology
KPI	Key Performance Indicator
MDM	Master Data Management
ML	Machine Learning
RDBMS	Relational Database Management System
PAAS	Platform as a Service
SAAS	Software as a Service
SLA	Service Level Agreement



We create impact that lasts by combining tech, data and human insight.

Solita is a community of highly and widely skilled experts geared for impact and customer value. We do what matters to build the future with our clients by delivering high-quality solutions to real problems.

Our unique service portfolio seamlessly combines expertise from strategic consulting to service design, software development, analytics and data science, cloud and integration services. Founded in 1996, Solita is a fast growing community of almost 1,000 professionals in Finland, Sweden, Denmark, Estonia, Belgium and Germany.

solita.fi/en | @SolitaOy