For next wave of innovation organisations will need internal data services

Abhishek Tiwari

Citation: *A. Tiwari*, "For next wave of innovation organisations will need internal data services", Abhishek Tiwari, 2016.
doi:10.59350/gdvr3-df924

Published on: April 22, 2016

To unlock the true value of data, organisations will need internal data services. Data services provide streamlined and centralised data access to a diverse set of users which removes the friction in delivering faster insights, products and services. Data services promote innovation. In addition, effective implementation of data services can reduce data duplication hence reducing the cost of data management and storage. Often referred as data-as-a-service (DaaS), data services are natural next step in the evolution of as-a-service model. Data services provide access to a broad range of internal and external data sources to power both internal business applications as well as client-facing software-as-a-service (SaaS) applications.

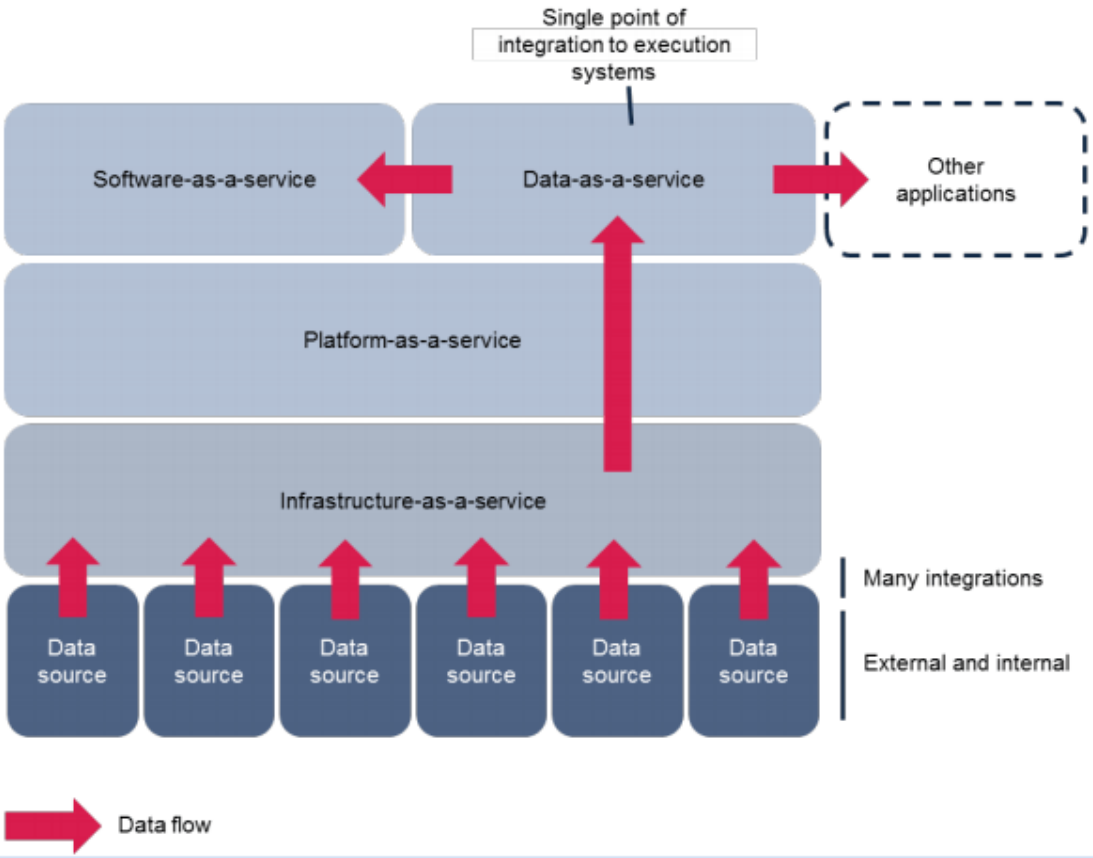

Figure 1: DaaS in the as-a-service stack. Credits Ovum.

A physiological analogy

If you can think a big and complex organisation as a human body then data is the oxygenated blood for the organisation. In a similar vein, you can view data services as a manifestation of the human circulatory system. Very much like human circulatory system which is responsible for the circulation of oxygenated blood in the human body, data services provide seamless data access to the whole

organisation.

More than just the technology solution

Vendors such as Oracle, IBM, SAP, etc. see data services as a natural extension of the as-a-service stack hence purely as the technology solution. There is no doubt that organisations will need right kind of technology solutions to build data services. Albeit I believe organisations should take a more pragmatic approach when building data services. Rather than a big bang approach, organisations should take a phased approach for building data services with each step creating a tangible benefit. A big bang approach has its own pros and cons but one of the biggest challenges is a 6-12 month wait period before the organisation can realise any benefits from a new data services regime.

One important aspect of data services is the skills. Organisations will need at least two types of skills - a data engineering team to build and operate the data services and a data science team to extract the actionable insights and prototype new products using the data services. Moreover, business users need to be educated on possible benefits both long run as well short term.

Another key aspect of the data services is the value realisation which requires monitoring and measurement demonstrable value. A mechanism to perform robust cost/benefit assessment of the data services on an ongoing basis is highly recommended. There are many ways to measure the value of data services but quantifying ROI can be incredibly difficult if not impossible.

Key characteristics of data services

Open Source

If you are not building your internal data services using open-source technologies then you are simply doing it wrong. If strong support and enterprise grade software are of paramount importance in your organisation then opt for a commercial distribution of open-source technologies. Rather than using open-source Apache Hadoop you have options to go with MapR, Hortonworks, Cloudera, etc. Similarly, you can use community edition of Apache Kafka or choose the commercial Confluent platform. With open-source technologies, there is always a right tool for the right job.

Canonical architecture for data services

Start with a canonical architecture for your internal data services. Identify data producers, data sources and data consumers in your ecosystem. Ensure that a service layer is acting as an interface to ingest and serve the data. Ingestion at scale can be supported using a modern message bus

(Pub/Sub), database change data capture or data ingestion framework such as Gobblin. Process the ingested data, in real-time using stream processing framework, batch processing or both - and store the processing results to appropriate data stores and databases for serving.

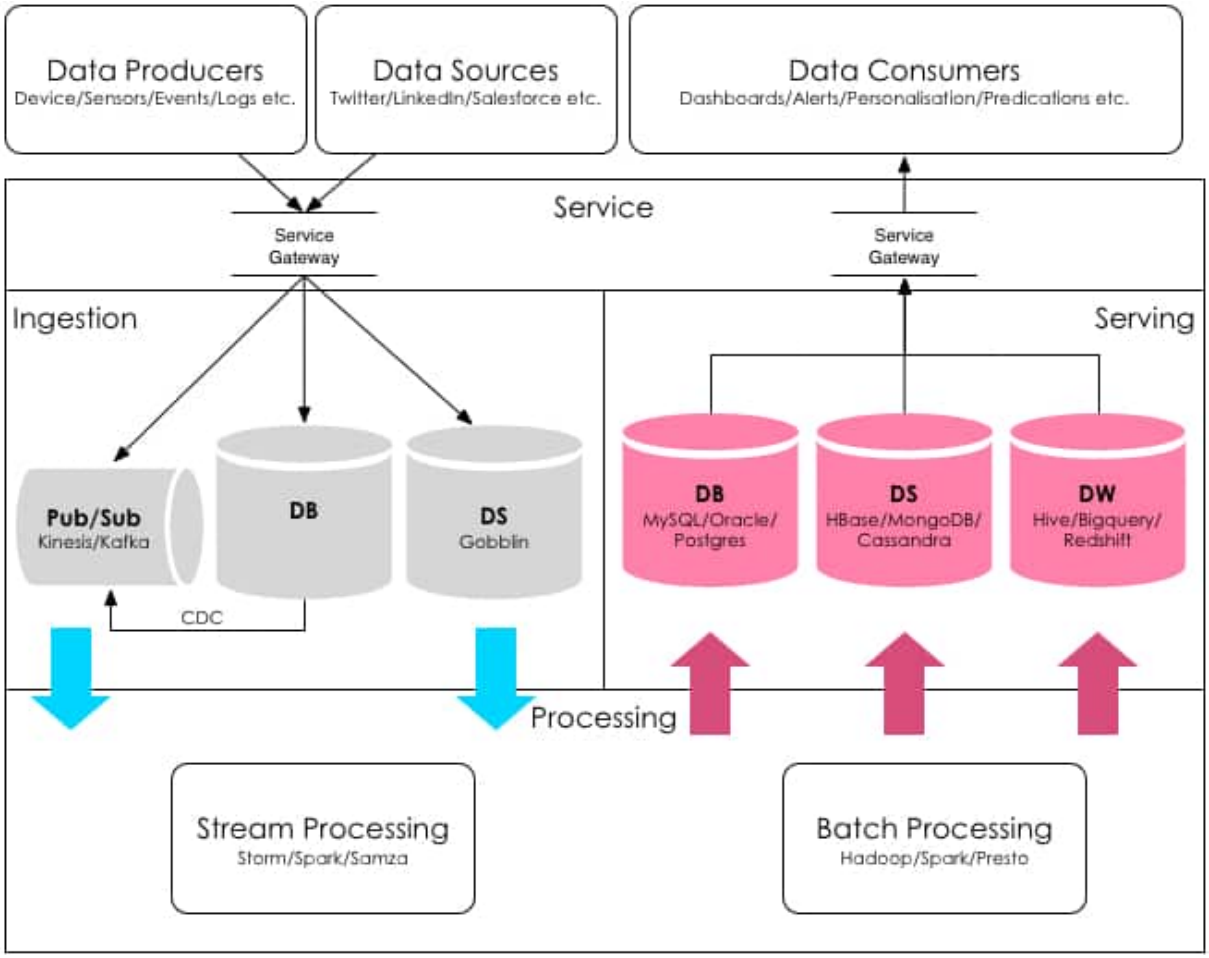

Figure 2: Canonical Architecture for data services

Plug-and-Play Architecture

A plug-and-play architecture is a must if you want to extract the maximum value out of data services. Both, data producers as well as data consumers should be able to tap into organisation's data services with less or minimum effort.

To enable the plug-and-play architecture a publish–subscribe pattern is a quite obvious choice. Technologies such as Apache Kafka and Amazon Kinesis can easily implement this pattern and highly recommended.

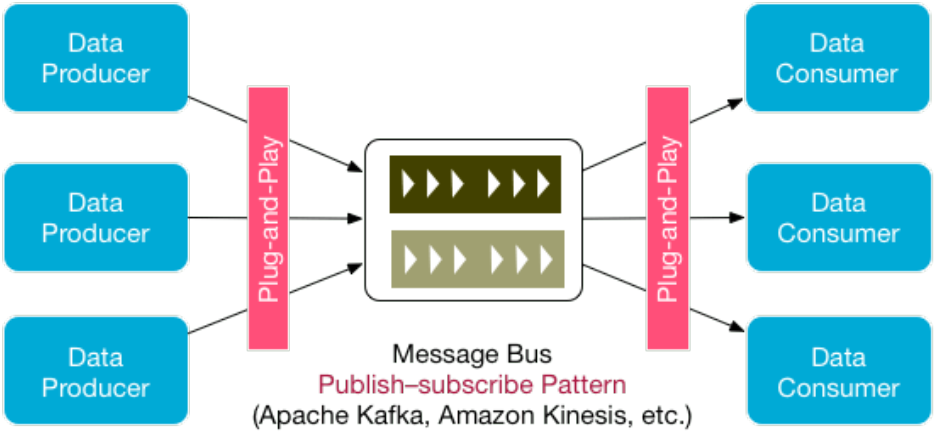

Figure 3: Plug-and-play architecture using publish–subscribe pattern

Data streaming and streaming analytics

Data streaming and streaming analytics will eventually replace traditional extract, transform and load (ETL) type of data analytics. Hence, design the data services for data streaming and streaming analytics. Especially if the organisation is dealing with real-time data and interested in data activation, streaming analytics performed on real-time data streams is highly valuable. A key benefit of data streaming and streaming analytics is real-time operational intelligence and KPI dashboards which can be useful for organisations across the board.

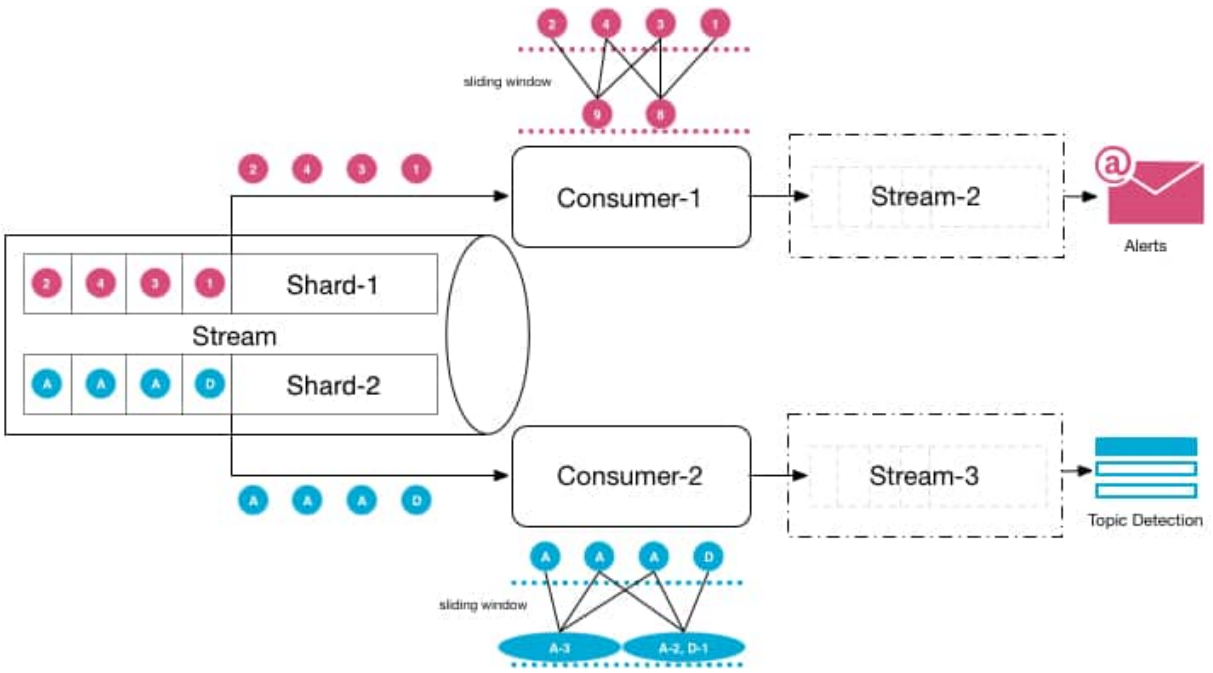

Figure 4: Data streaming and streaming analytics

Cloud-native architecture

To support cost-effective and scalable on-demand data service use of cloud-native architecture is highly desirable. When we talk about the cloud-native architecture we are referring concepts such as stateless architecture, single point of failure (SPOF), data partitioning, etc. Public cloud or hybrid cloud, implementation of cloud-native architecture decides success and scalability of data services.

Data lake and data warehouse

As part of data services, organisations require a data lake and optionally one or more data warehouses. A data lake backed on object-based storage repository such as Amazon S3 can hold data in its raw format until it is needed. For batch analytics, data lake and data warehouse are quite essential and compliments the streaming analytics.

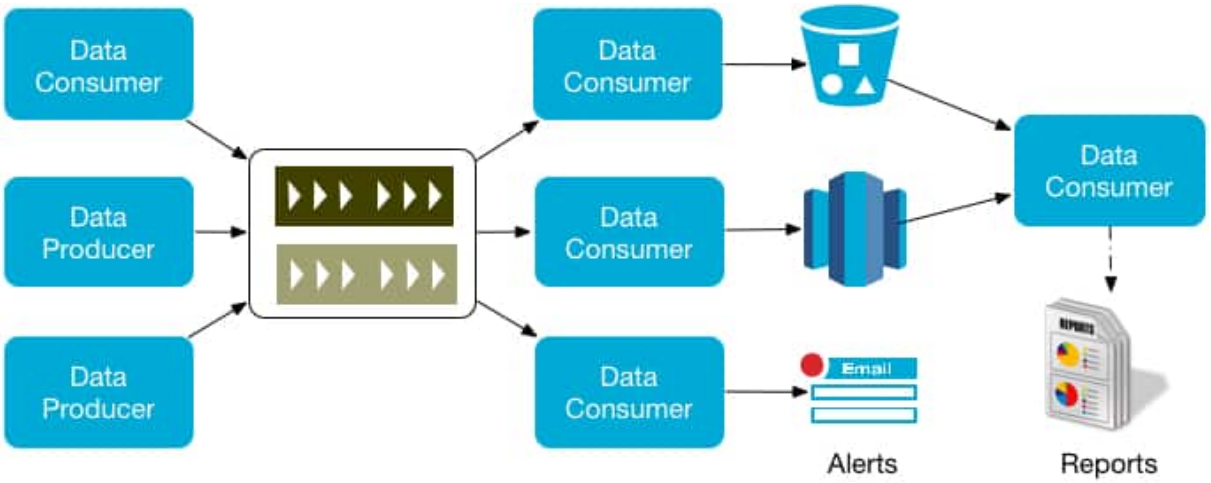

Figure 5: Data lake and data warehouse

Data governance, privacy and security

For most of organisation data volume and variety is continuing to expand, hence making data governance, privacy and security more difficult than ever. Organisations need to implement clear robust policies around data access, data lifecycle, data archiving and data deletion. The implementation of data access is abstracted so that it separates data producers and consumers from the data stream and data repositories.

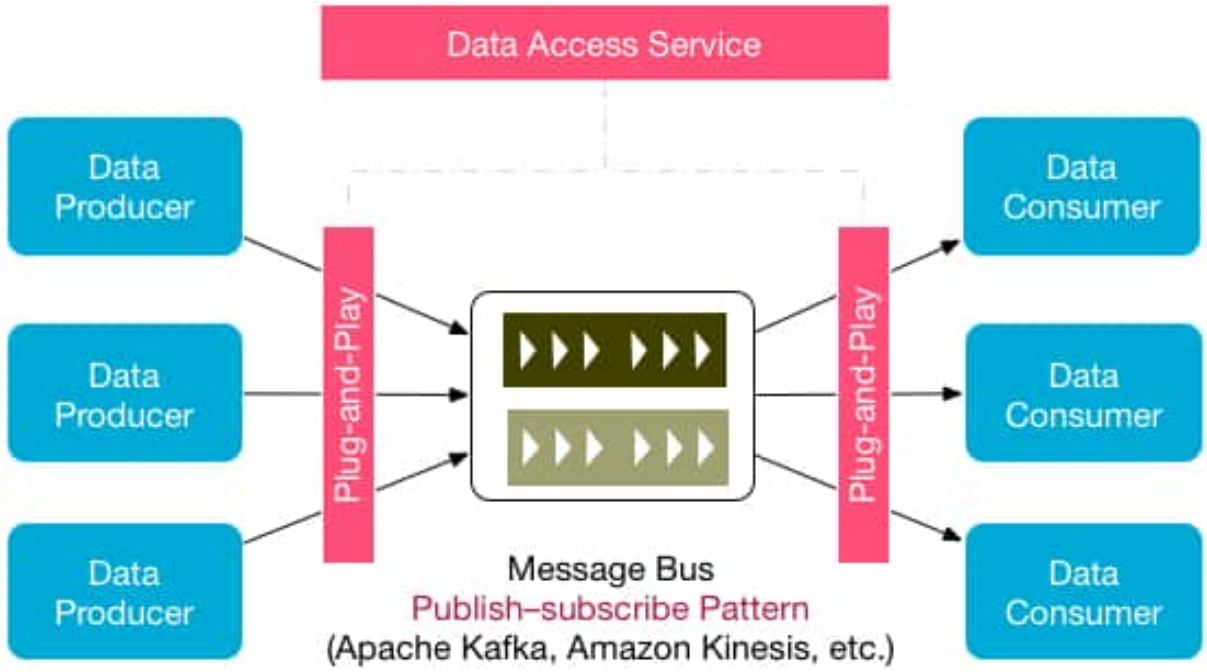

Figure 6: Abstracted data access service

Extensible authentication, granular access controls, and comprehensive audit logs are must if you want to establish a strong data governance and security framework. These need to be implemented at an individual component level as well as solution level - with seamless interoperability using something similar to single sign-on. For example, single authentication should be used to access a Redshift Datawarehouse and a MongoDB cluster but with varied access level as defined by components specific Access Control Lists (ACLs) or Access Control Expressions (ACEs).

Last but not least, in a highly-regulated environment such as banking or insurance data lineage i.e. ability to track back the source of data and applied changes to identify how the final data sets arose is often a quintessential requirement. Using comprehensive audit logging at each and every data manipulation stage coupled with a unique data identifier one can track down whole data lineage.