How to better serve and predict your digital business

With fast and streaming data

TRIVENTO

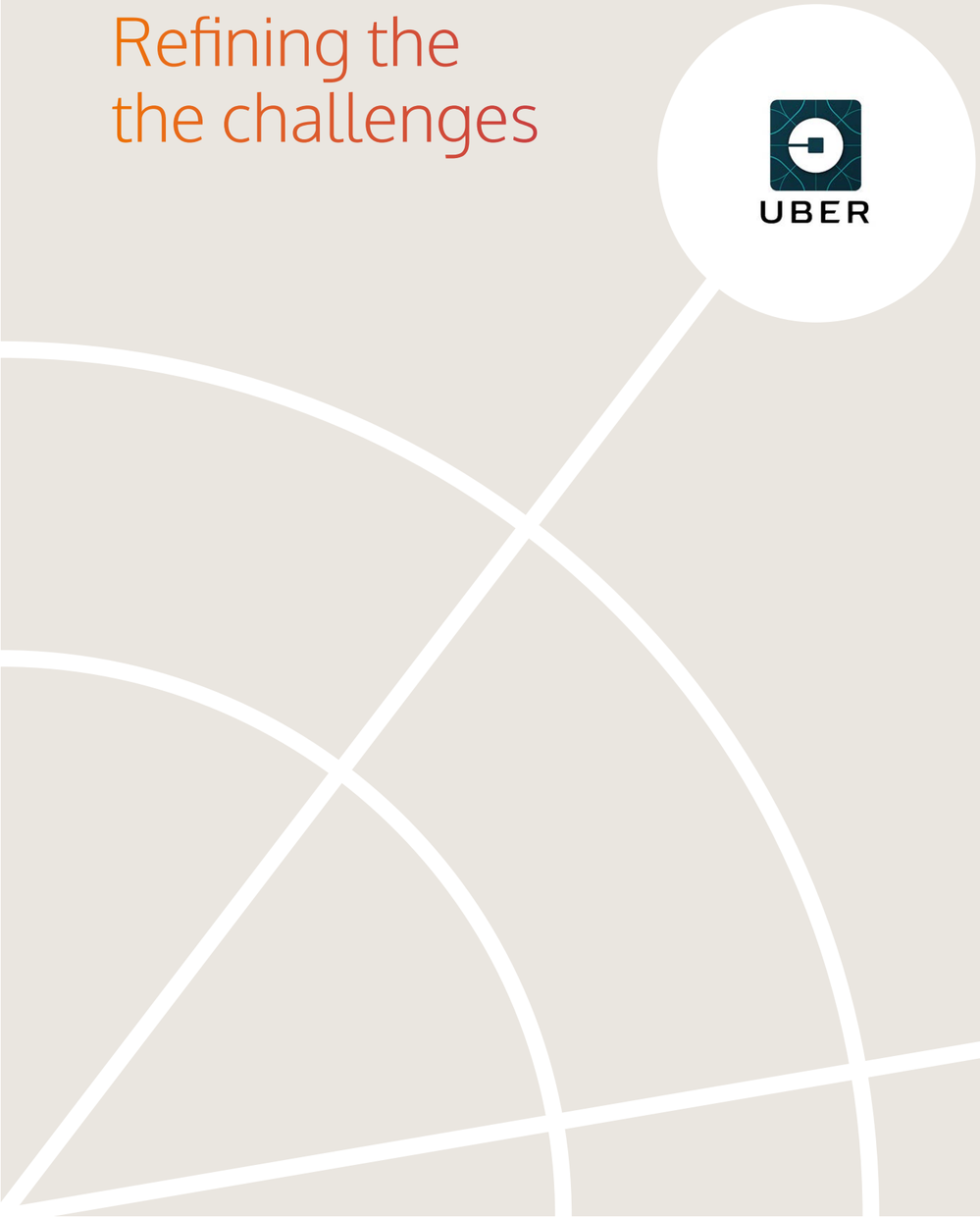
Decision makers increasingly require data in realtime. But how does realtime data improve your digital business? Why is a realtime data & analytics platform a crucial enterprise building block? How can it be implemented in an Agile way?

In this whitepaper you will find more insight with regard to needs, challenges, use cases and possible solutions to make money out of fast or streaming data. But first the question, what is fast or streaming data exactly?

Fast data is often defined as data in motion, so called streaming data. This comes from hundreds, thousands or maybe even millions of endpoints – mobile devices, web apps, sensors, financial transactions, logs, authorization systems and more.

It results in streaming data that can be used to make realtime decisions, to improve efficiency and lower cost in operations, maintenance and commerce. Streaming data can be seen as an additional datastream to big data.

Refining the the challenges



Uber needs to present the user an interesting offer in milliseconds and to millions of users all over the world.

In order to do this Ubers realtime matching service uses fast data technology to match any request from anywhere and anybody in the world, with available cab drivers, including route costs for different route/driver options.

This proces of returning results in an offer that is sent to a customers mobile app. An offer is needed under a second in order to meet customer demands and expectati-ons.

E-commerce websites require real-time offers and recommendations, to be redeemed instantly.

Therefore you need to be able to match customer information, stock and discounting policy information against the ability to redeem it in realtime.

This requires to be fast data and the technology that is able to support milliseconds roundtrips from request to respond. Why is this important?

54%

54% of the respondents stated that leveraging fast data becomes more important than leveraging big data.

The speed at which organizations can convert data into action insight and is considered more and more critical depending on use cases. In market research by Capgemini and EMC, more than 50% of the respondents stated that leveraging fast data becomes more important than leveraging big data.

'Fast data becomes more important than big data as people expect realtime, personal and relevant information from a realtime bank.' ***ING Bank - survey State of Fast Data Trivento 2017***

Execution times for handling data evolves increasingly becomes faster reducing the time ot takes from from hours and minutes to microseconds.

In the figure on the right you can see that typical execution times for handling data evolves from hours and minutes to microseconds. This also counts for times between data collection and processing.

Systems and applications that are designed to take advantage of Fast Data, allow companies to make realtime decisions with interactions.

Fast data systems or platforms operationalize the learning and insights that companies take from big data.

Metric	Sizes and units: Batch	Sizes and units: Streaming
Data sizes per job	TP to PB	MB to TB (in flight)
Time between data arrival and processing	Many minutes to hours	Microseconds to minutes
Job execution times	Minutes to hours	Microseconds to minutes

What is the relation with Big Data?

Where fast data is often defined as data in motion, Big Data is static; also called data at rest.
Big Data describes data's volume - petabytes to exabytes - and their variety: structured, semi-structure and unstructured data.

Big Data systems facilitate the exploration and analysis of stored, large data sets. The total of all stored data is often called a data lake.

Once data is no longer interactive and fast moving, it will be added to the data lake, whose responsibility it is to provide reliable, scalable storage and tools to query this historical data in the future.

What are the use cases that require fast data solutions?

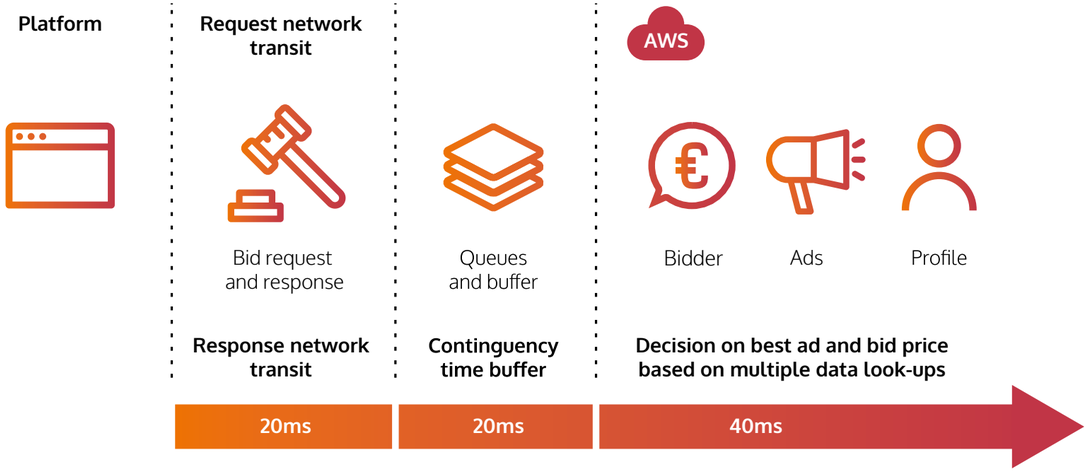
In addition to the Uber case we described earlier, the proces of realtime bidding and smart metering in terms of Internet of Things (IoT), also requires fast data solutions. These examples are to help you determine if your challenges can be solved with a fast data solution or require a different answer.

Realtime bidding

When you visit a website you are served a wide range of advertisements. The process to serve you that ad involves the website or publisher contacting an ad-exchange, which then accepts realtime bids from many different parties and returns a matching advert.

In fact the use case requirements are quite similar to those of the use case realtime recommendations. With the difference that the served content (advertisements) are coming from different bidders that are involved, potentially on a global scale.

The bidders can do an offer for the serve of a commercial banner, based on the available information about that user (website, ad location and size, demographic information, such as user location, browsing history, and time of day). The price they want to pay is based on the available information about the user and it's relavancy.



This data (content) may be provided by publishers (most often the case for mobile applications or websites) or third-party data providers. Whichever bidder bids the highest within a fixed time period set by the exchange gets to serve the ad and pay the bid price. This usually under 100 milliseconds.

Smart metering

Smart metering (IoT) is another example of the benefits that can be achieved with fast data. There are significant asset management challenges that come with the management of physical assets. Streaming or fast data architectures are required in order to be able to

manage the data that is produced by thousands of sensors attached to assets (electronic engines, smart meters) that are in use at any given time. New software architectures and technology are needed as the internet is now used as a default medium to transport these data streams.

To realize business value, all this data needs to be interpreted and turned into meaningfull information. The first challenge is to ingest streams of data coming from this big number of sensors. Imagine ingesting on average 10 readings a second from 100,000 devices. This represents a large ingestion stream of 1M events per second!

But ingesting this data is only the smallest and simplest of the tasks required in managing these types of streams. As sensor readings are ingested, a number of decisions must be made in realtime. This means against each event, as soon as the event arrives.

Are all devices reporting data as they are expected to? Are any sensors reporting values that are outside the defined bandwidth? This demands a constant live monitoring of the data as it is received.

Utility sensor readings become much more informative and valuable when one can compare a reading from a single meter to 10 others, all connected to the same transformer. This makes it possible to determine if there is a problem with the transformer or with a single meter located at a home. Toon, the smart energy monitor is a prime example of this behaviour.

Emilio Martinez de La Fuente

Head of Global productmarketing
Smart Metering & Grid, Telefonica

"Smart Metering greatest challenge is to design a comprehensive data management model that allows:

- data exchange flows among all stakeholders
- hosting huge data volumes
- secure the data
- valuable data analytics."

What is the business case?

To understand the needs for fast data within your enterprise, ask yourself: 'Why do organizations perform analytics in the first place?'

The answer lies in the need for businesses to make realtime decisions faster, with regard to use cases related to: fraud detection, customer engagement, online conversion, alerting and interaction.

The interactions that are involved, are provided by responsibility of the web and mobile apps, and the most valuable improvements come when these interactions are specific to the context of each event and in realtime.

Returning to the smart meter example, transformers show a specific trend before they fail. Failure in service or it breaking down could thus be prevented.

What makes it important to identify these failures before they happen? Streaming data can help identify failures and reroute power to alternative transformers.

'With the help of streaming analytics we are able to predict the effect of changing conditions on the behaviour of our systems' ***IoT Champion, Pentair.***

Businesses want to be in control and avoid unnecessary disruption of processes. Now that smart, internet connected, sensors are being embedded into more and more devices, companies can use the data these sensors provide to determine when maintenance is required for example in planes, trains, machines, and household appliances in a preventive way.

Ideally, when an issue has arisen, companies want to understand the issue, what caused it, and how it can be resolved. Preferably before a maintenance professional or crew has been dispatched.

In all use cases the availability of realtime data in the operational process leads to the ability to support decisions that result in better conversion rates, higher sales per transaction, lower cost of production failures.

What if BP had detected problems before the oil hit the water?

What does a typical fast data solution look like?

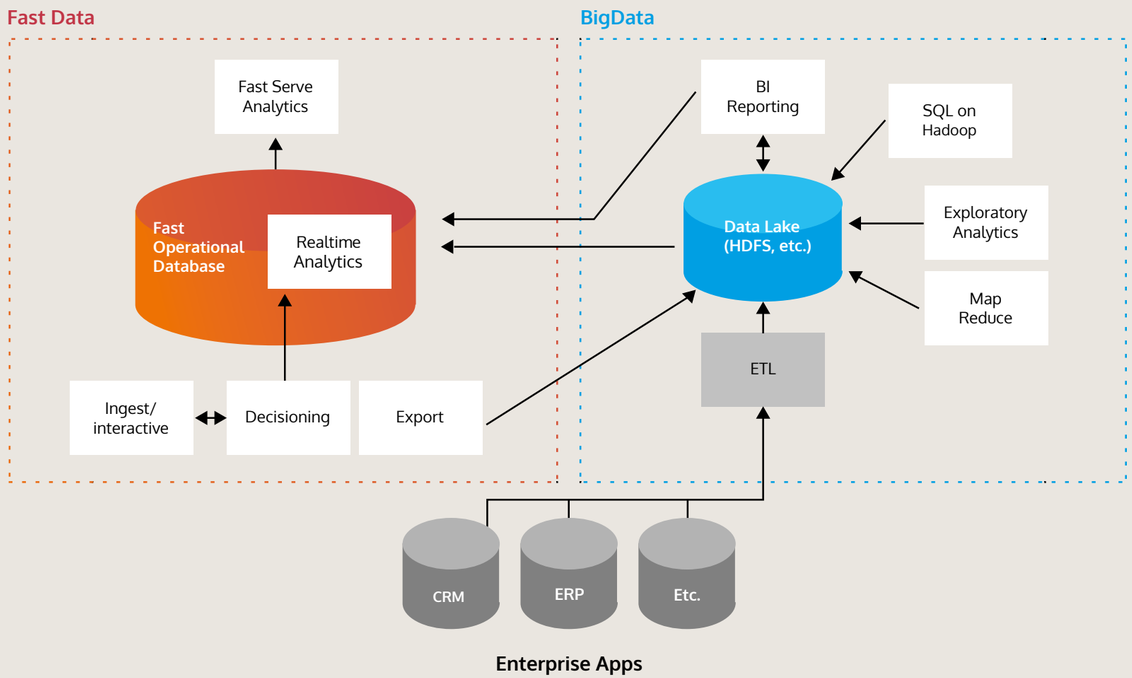
When we take a look at the logical solution needed for fast data use cases, it is most often positioned in relation to an existing big data architecture.

Streaming data requirements lead to an additional architecture and software stack. This is necessary in order to be able to implement fast data use cases.

All data that is collected for realtime use in operational processes, finally ends up in the existing data lake. But before that, the data already adds value in supporting operational processes.

We already mentioned the added value of making better decisions supported by realtime data. The ability to serve the right content and valuable recommendations. This will result in the prevention of potential loss of customers, production time and valuable assets.

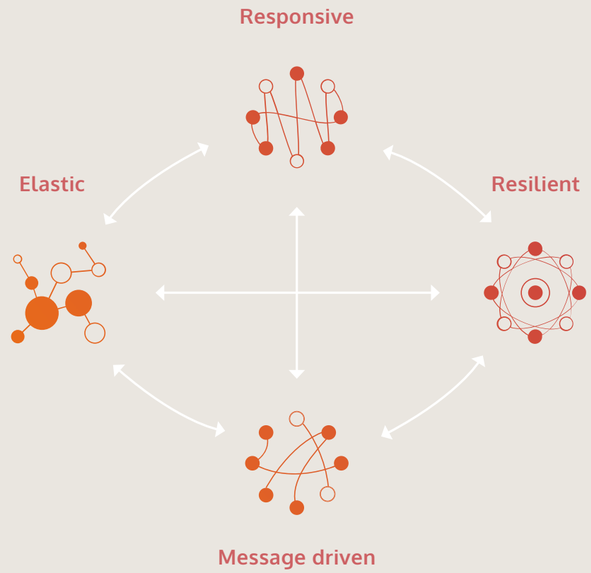
The combination of a big data and fast data architecture is often referred to as a lambda architecture. This also leads to the conclusion that the tools and techniques that are chosen for implementing a data lake, are simply not engineered to support fast data uses cases. Fast or streaming data use cases demand a specific fast data architecture.



The basic issue with streaming data solutions is the need for distributed computing power and software to handle high volumes of data in very short times. This need can be provided by a cloud infrastructure in combination with a distributed software architecture.

Fast data architectures can be realized by means of distributed data processing. Batch jobs last for hours but a streaming pipeline is designed to run for weeks, months or even years. A different approach is needed to guarantee its availability.

The required qualities can be expressed by the reactive paradigm.



Responsive

The business value is a consistently responsive user experience that is highly available, never fails during busy times, and is not susceptible to blocked processes and cascading failures.

Not all components of the system need to be resilient. The system as a whole however needs to be responsive. A microservices architecture will help you realize responsiveness in the business critical parts of the system.

Resilient

To achieve responsiveness, the system needs to be resilient against failure of any major issue, server crashes, hard drive failures or network partitions loss.

Replication of data (distribution) prevents data loss and enables (micro)services to keep working. Designing isolated functions with a bounded context in single microservices will prevent the cascading of failures.

Asynchronous, non-blocking, message-driven communication enables supervision, isolation and replication of failed processes.

Elastic

Scaling out or up across all physical and cloud infrastructure during busy times, and lowering costs by dynamically scaling in or down during slow times—is made possible through the level of indirection and loose coupling the messagedriven foundation enables.

Message driven

Services respond to messages: directed commands and queries. Furthermore, they use messages to send commands and queries to other services as well. Asynchronous, non-blocking, message-driven communication enables supervision, isolation and replication of failed processes.

How about the Data Lake?

The Data Lake can be seen as the Big Data part of the architecture. This is the central storage location for all of the enterprise data. The Data Lake is a critical attribute for a data pipeline.

The data lake is not necessarily unique because of its design or functionality. Its importance comes from being an enormously cost-effective system to store large volumes of data.

Essentially, it is a distributed file system on cheap commodity hardware. Hadoop Distributed File System (HDFS) is used a lot as a data lake solution, but not the only option.

Where do I start as a company?

So far the message is clear, but where and how do I get started with my fast data solution, might be the next question. To implement a fast data solution there are three basic steps that need to be taken:

1

Interact with the ingested data feed

Ingest data from multiple sources that deliver high volume data on a constant base. Inform actors, facing this data stream, on what to do with it. This could be a recommendation or alert.

2

Make context based decisions on each event in the feed

Combine realtime data with data from a data lake to allow for realtime decisions on how to respond to an event. This offers opportunities to enhance the online interaction. Additional (historical) data provides the context that is needed for making realtime decisions or doing realtime recommendations.

3

Provide visibility though realtime analytics

Run analytics against a fast data engine to inform operational decisions. The ability to take more than a single event into account when making a decision makes that decision that much more solid.

How we can help

We can help you implement, experiment and learn in several ways. Trivento is Lightbend consulting partner and partner of Amazon Web Services (AWS).

10 day Quickstart with Lightbend and Trivento

TRIVENTO Lightbend

Trivento and Lightbend team up to give you a headstart

Together with Lightbend we're rolling out the Fast Data Quickstart service to help your team get started right away. This 10 day service engagement ensures you your team is using the proper streaming technology for the right use case. Avoid rookie mistakes and get your project off on the right foot. Call now and make an appointment.

Trivento is Lightbend consulting partner
and partner of Amazon Web Services (AWS)

Alex van Koutrik
alex.van.koutrik@trivento.nl
06 - 46 14 94 00

Sources: Capgemini Big & FastData. The Rise of Insight-Driven Business, Amazon Web Services – Building a realtime Bidding Platform on AWS, VoltDB Big Data versus Fast Data, OReilly, Fast Data and the new Data Enterprise Architecture, Lightbend - Fast Data Platform, Data Lakes - Risk Management, Casper Tribler