Modern Data Architecture

Alexey Grishchenko

Pivotal

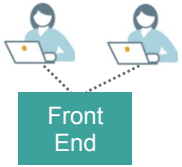
Pivotal Confidential–Internal Use Only

About me

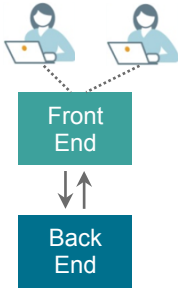
Enterprise Architect @ Pivotal

- 7 years in data processing
- 5 years with MPP
- 4 years with Hadoop
- Spark contributor
- http://0x0fff.com

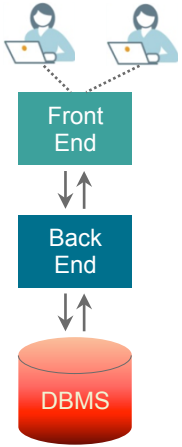
How it started…



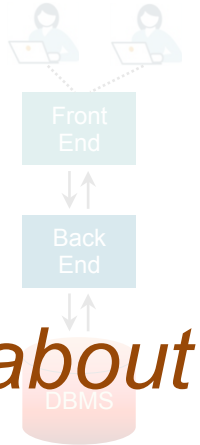
How it started…



How it started…

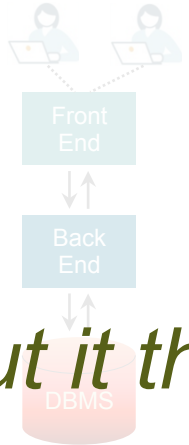


How it started…



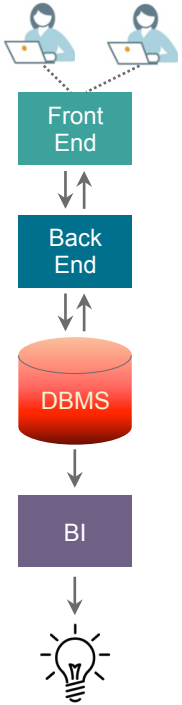
What about BI?

How it started…

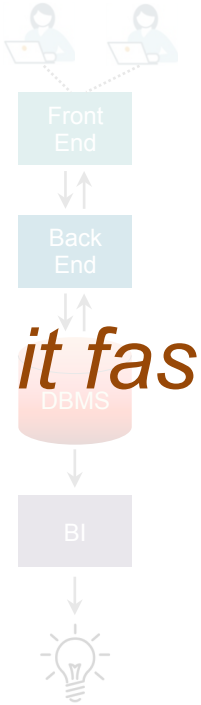


Just put it there!

How it started…

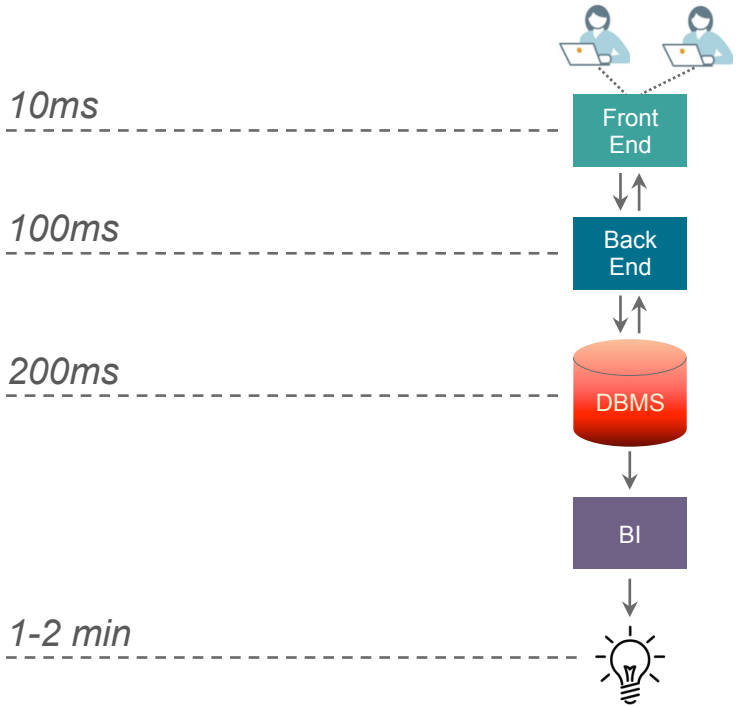


How it started…

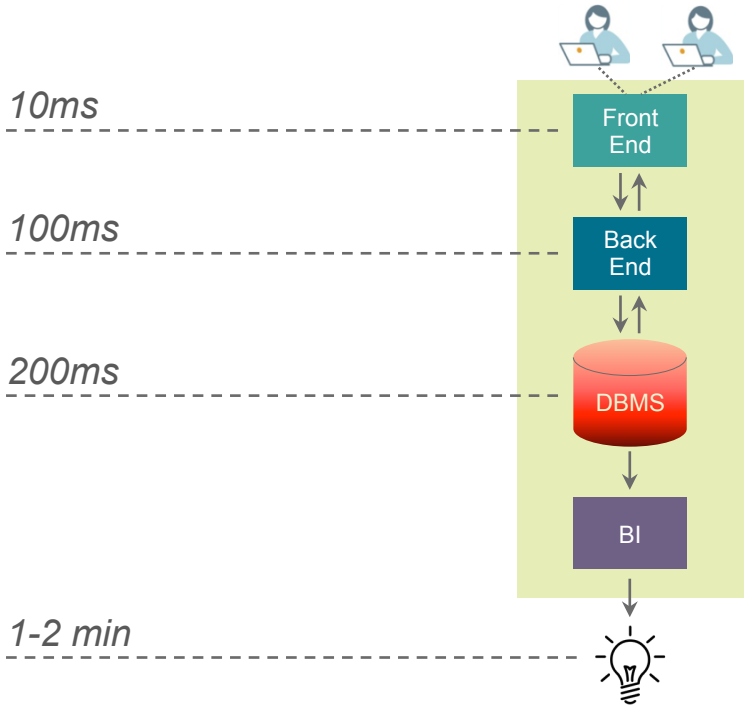


Was it fast?

How it started…

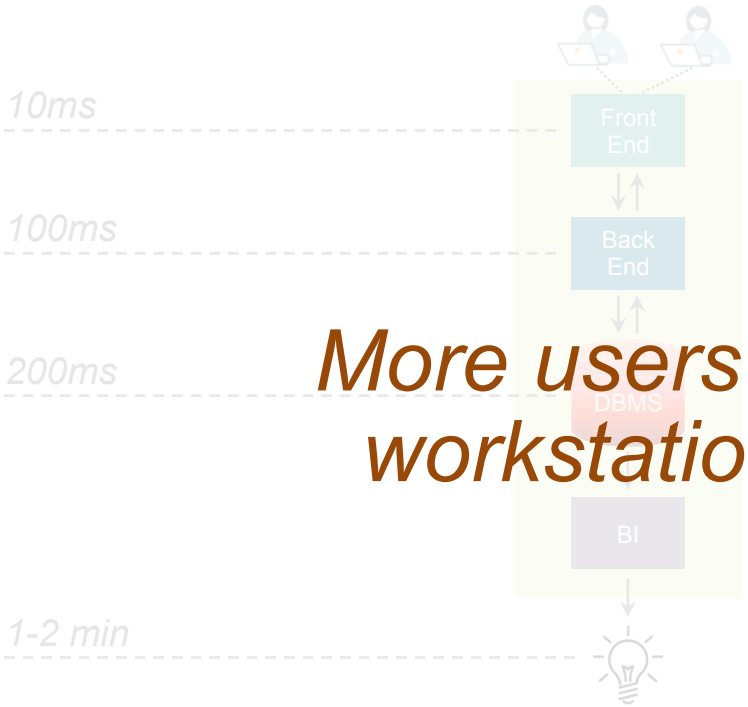


How it started…



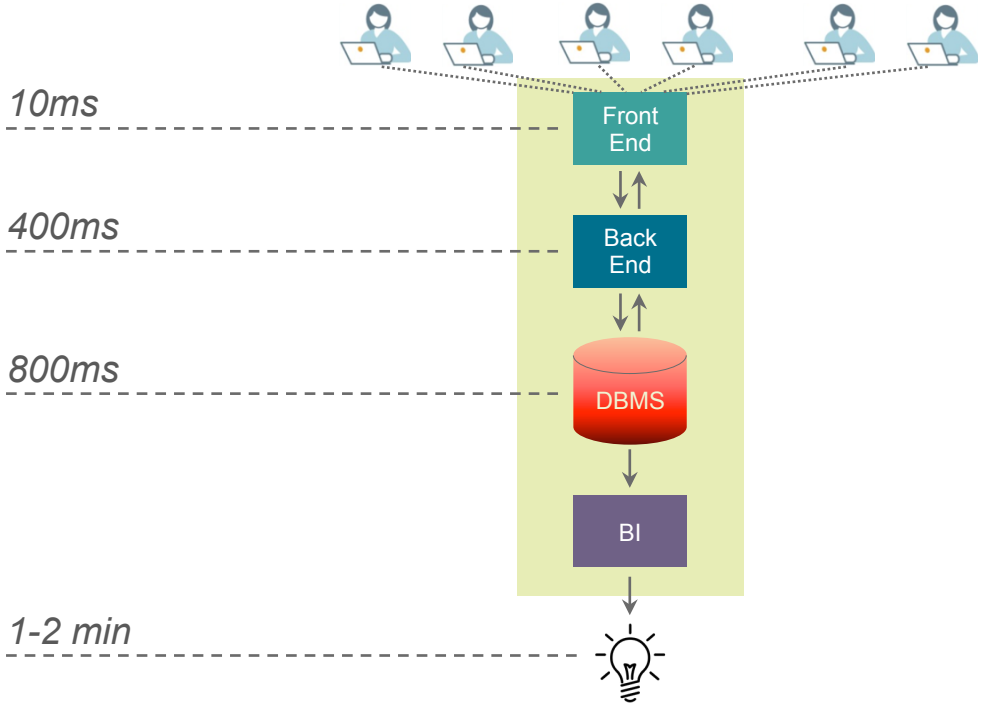
yes, single server…

First Issues

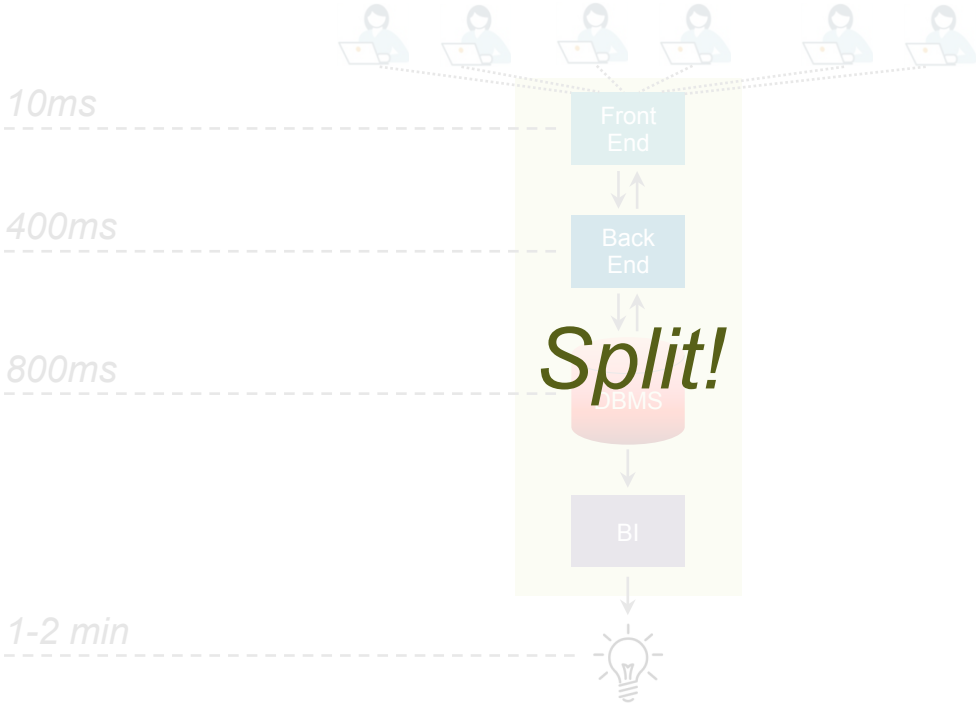


More users got workstations

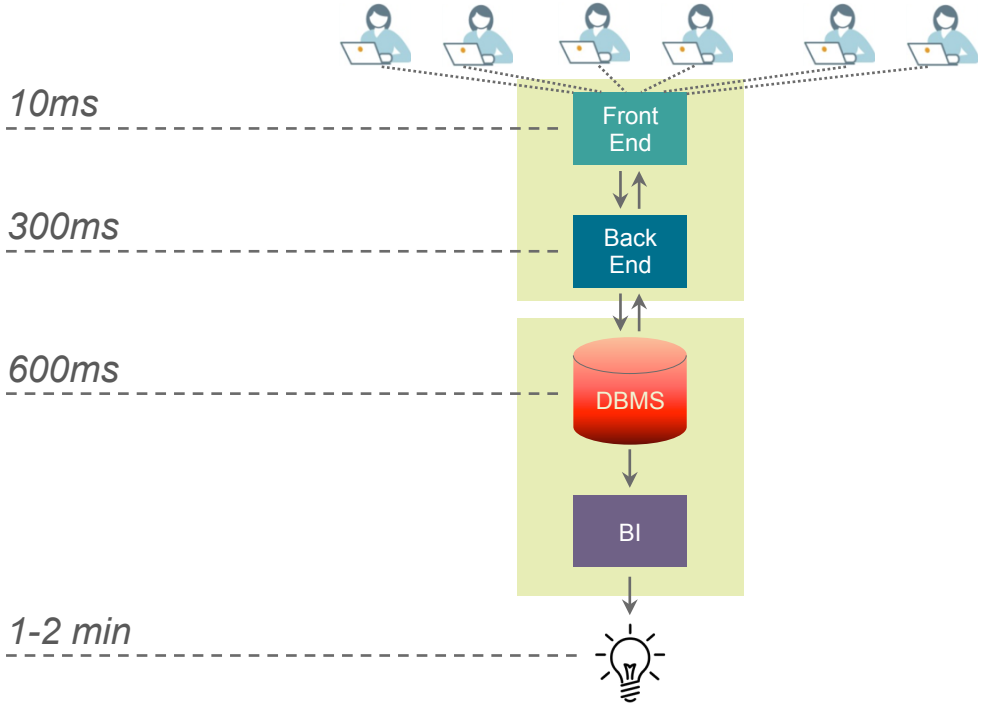
First Issues



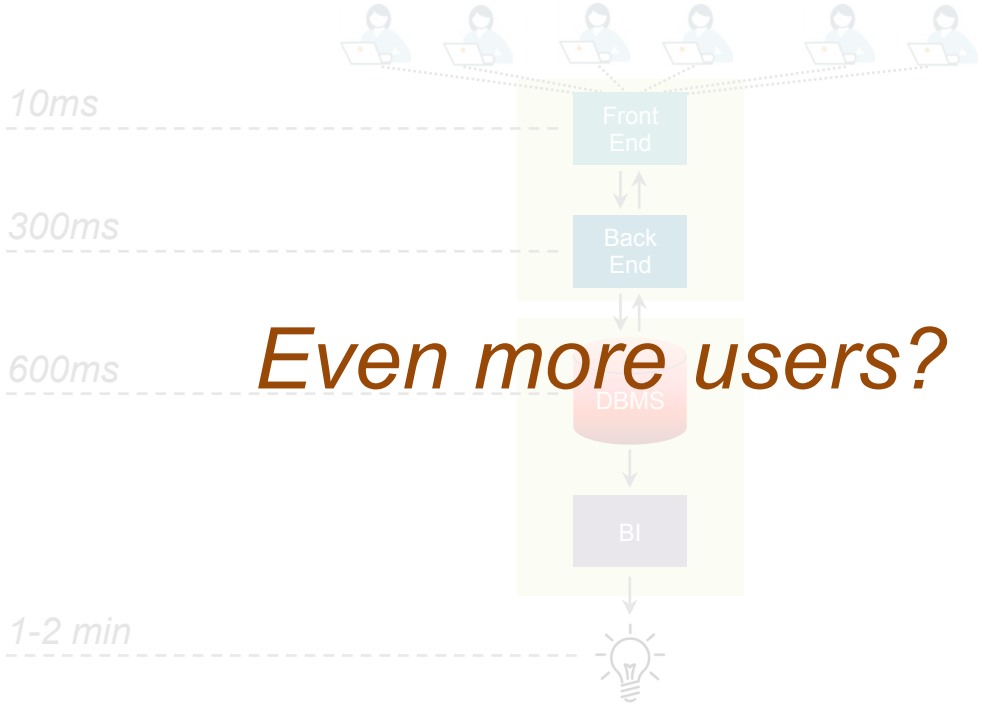
First Issues



First Issues

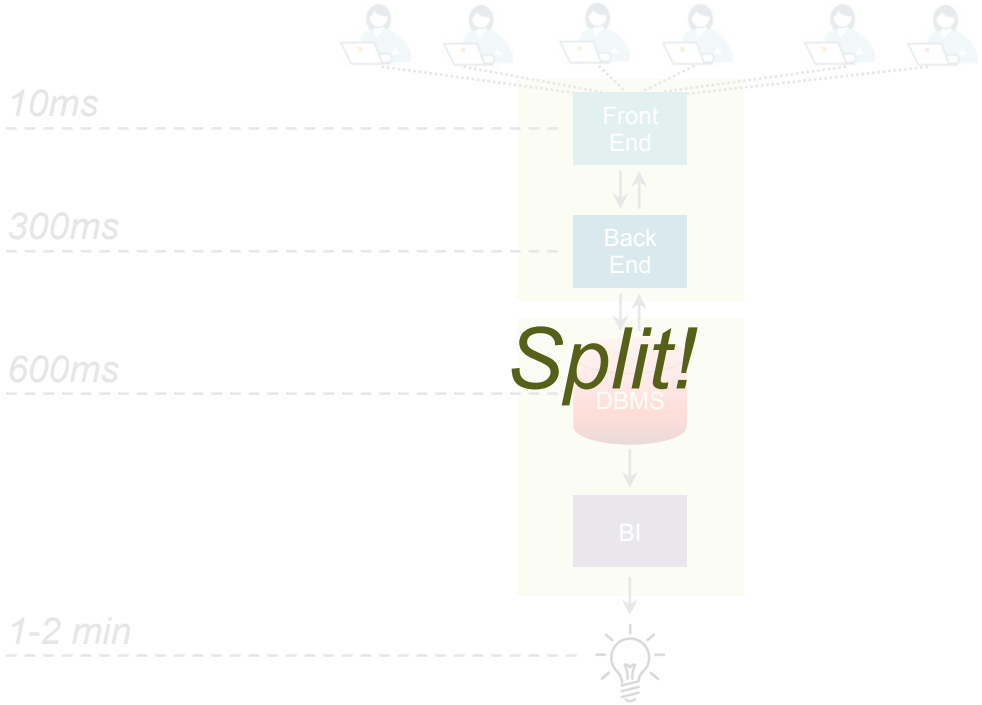


First Issues

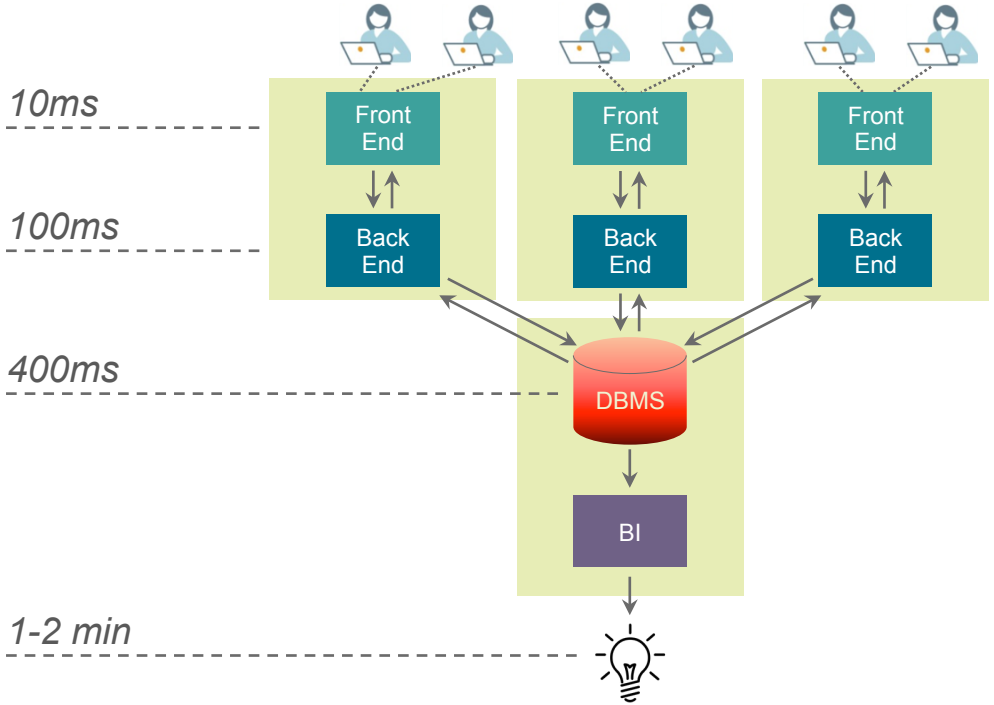


Even more users?

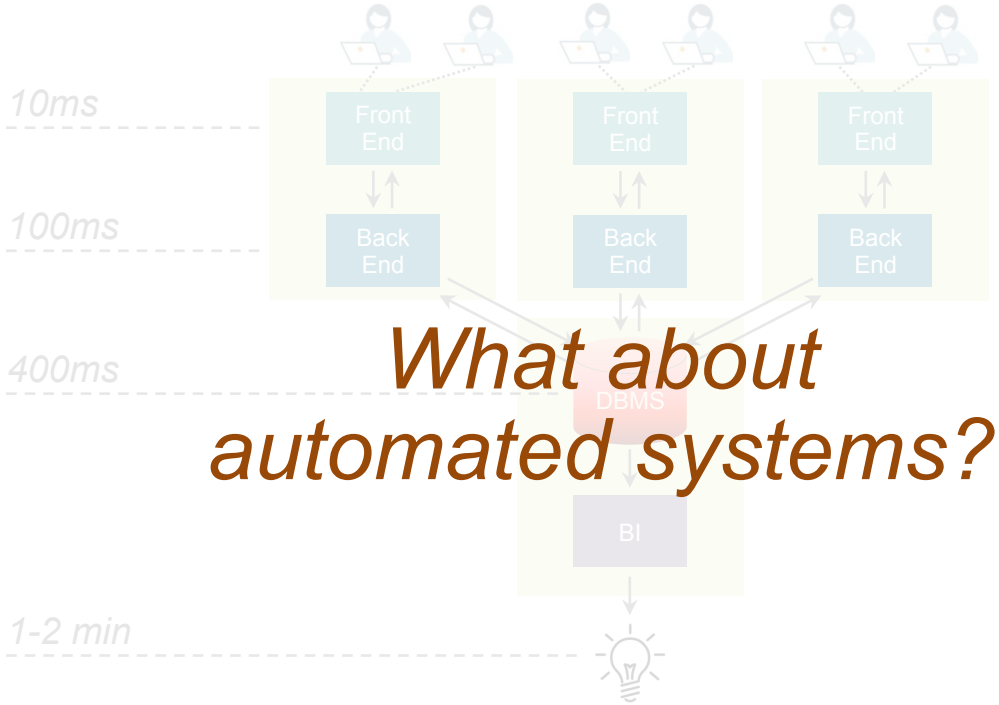
First Issues



First Issues

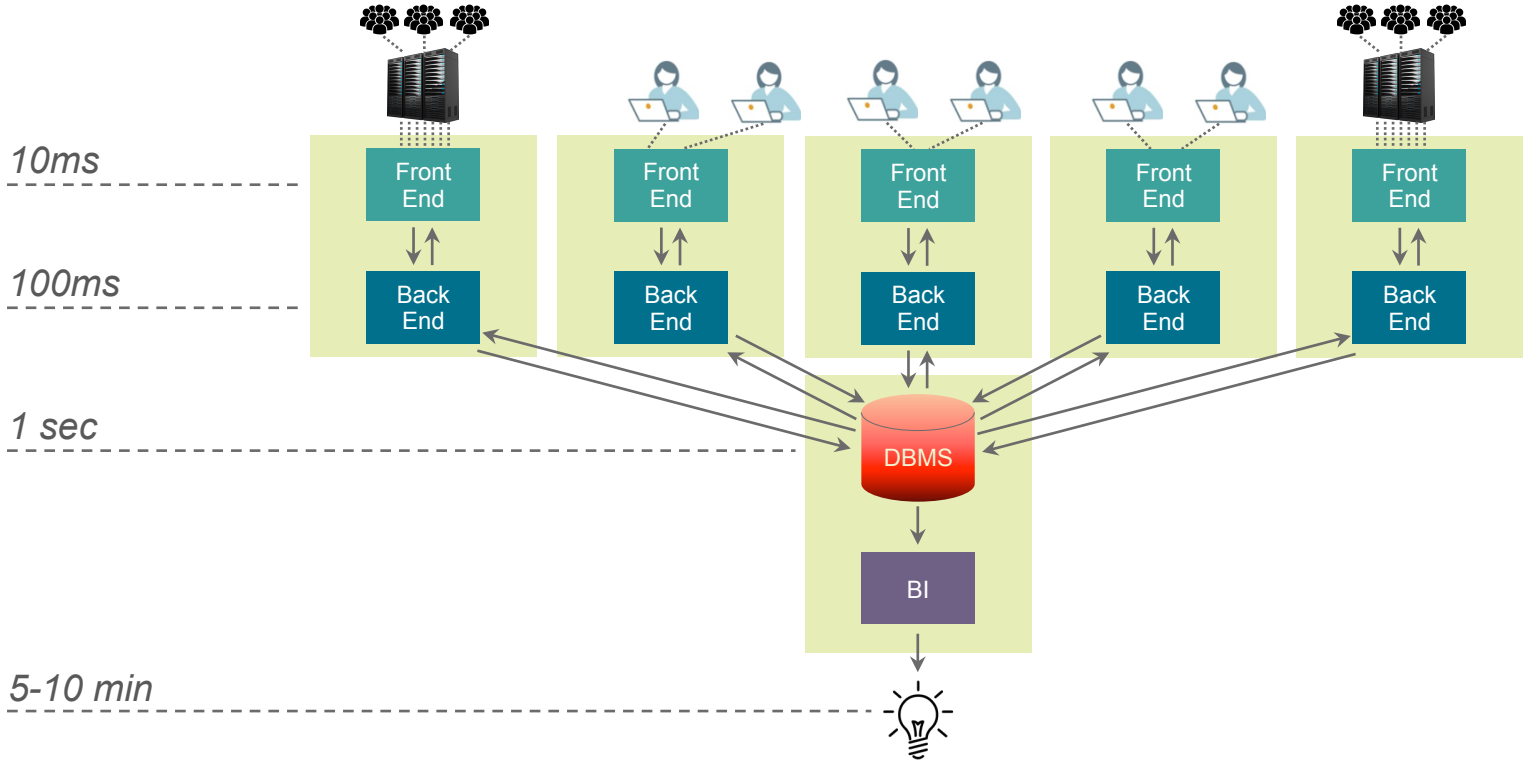


First Issues



What about automated systems?

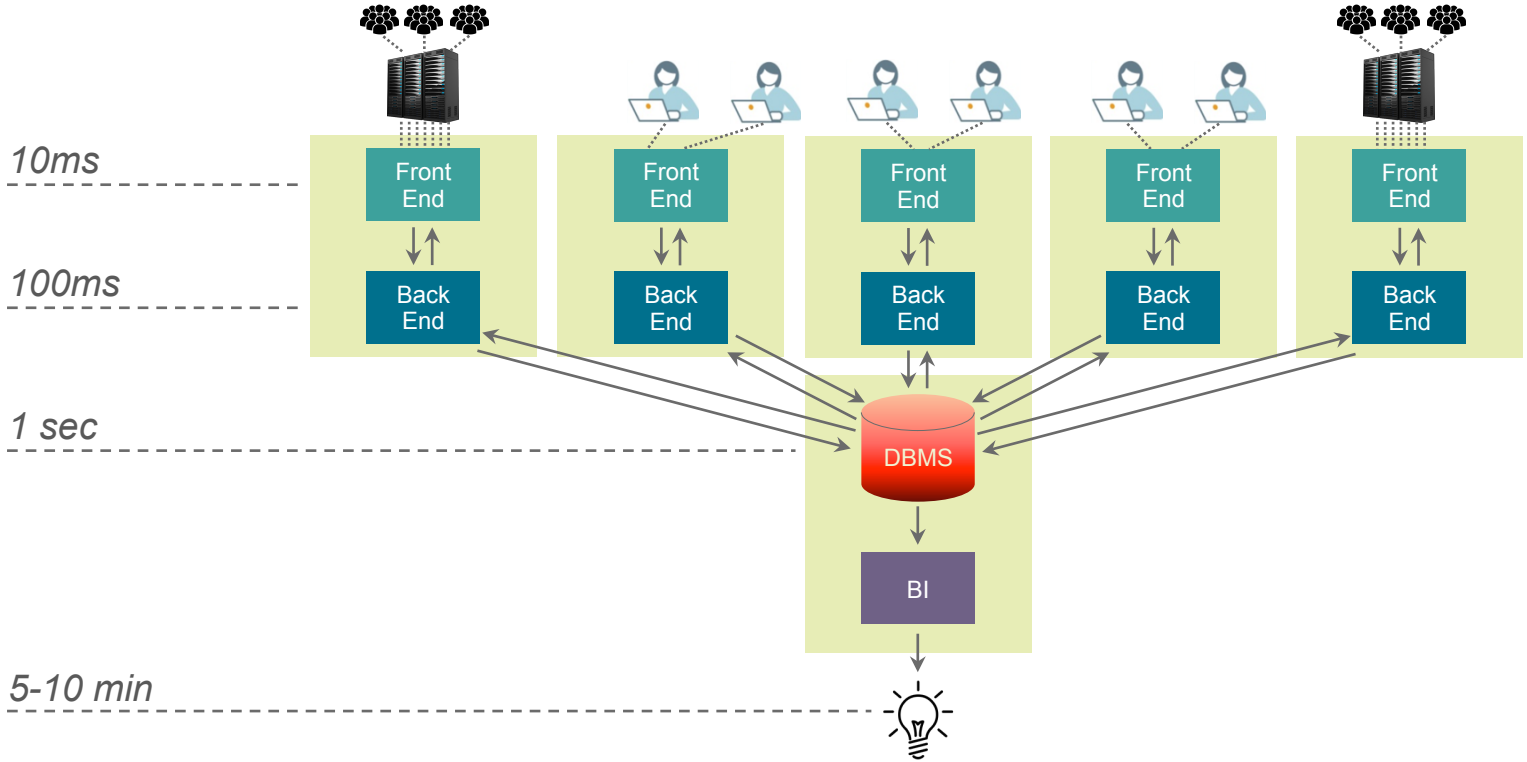
First Issues



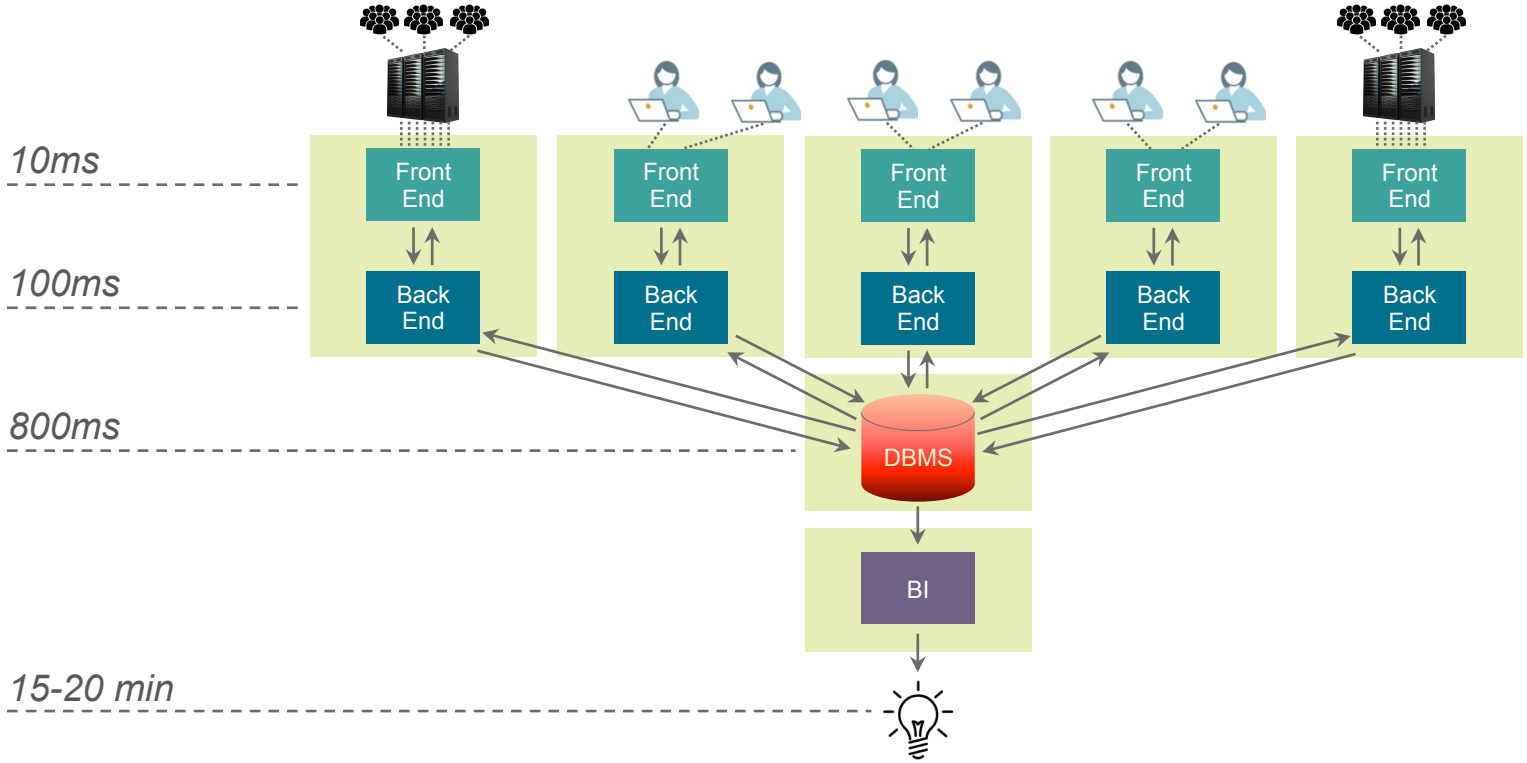
First Issues

Database, please, live!

First Issues



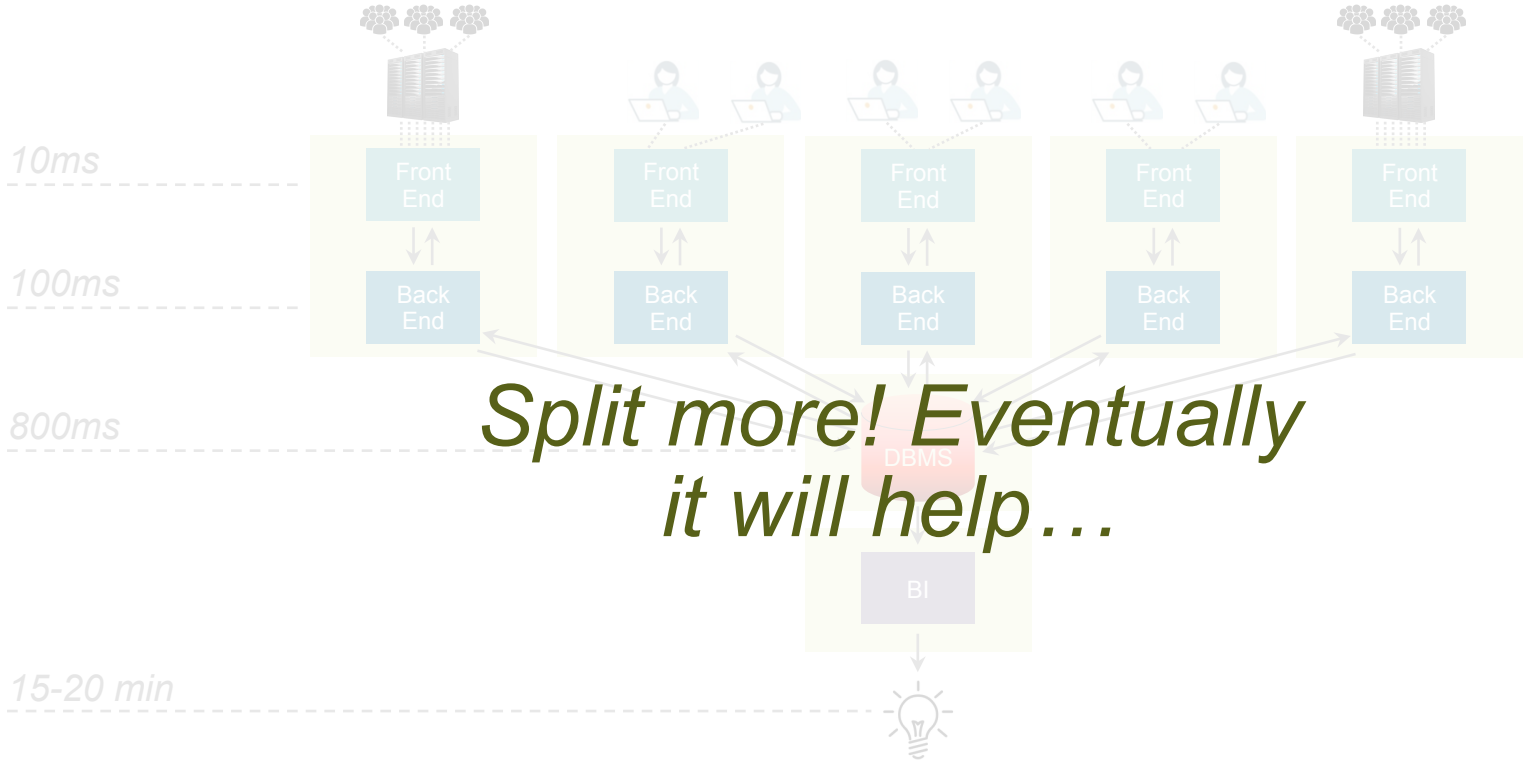
First Issues



First Issues

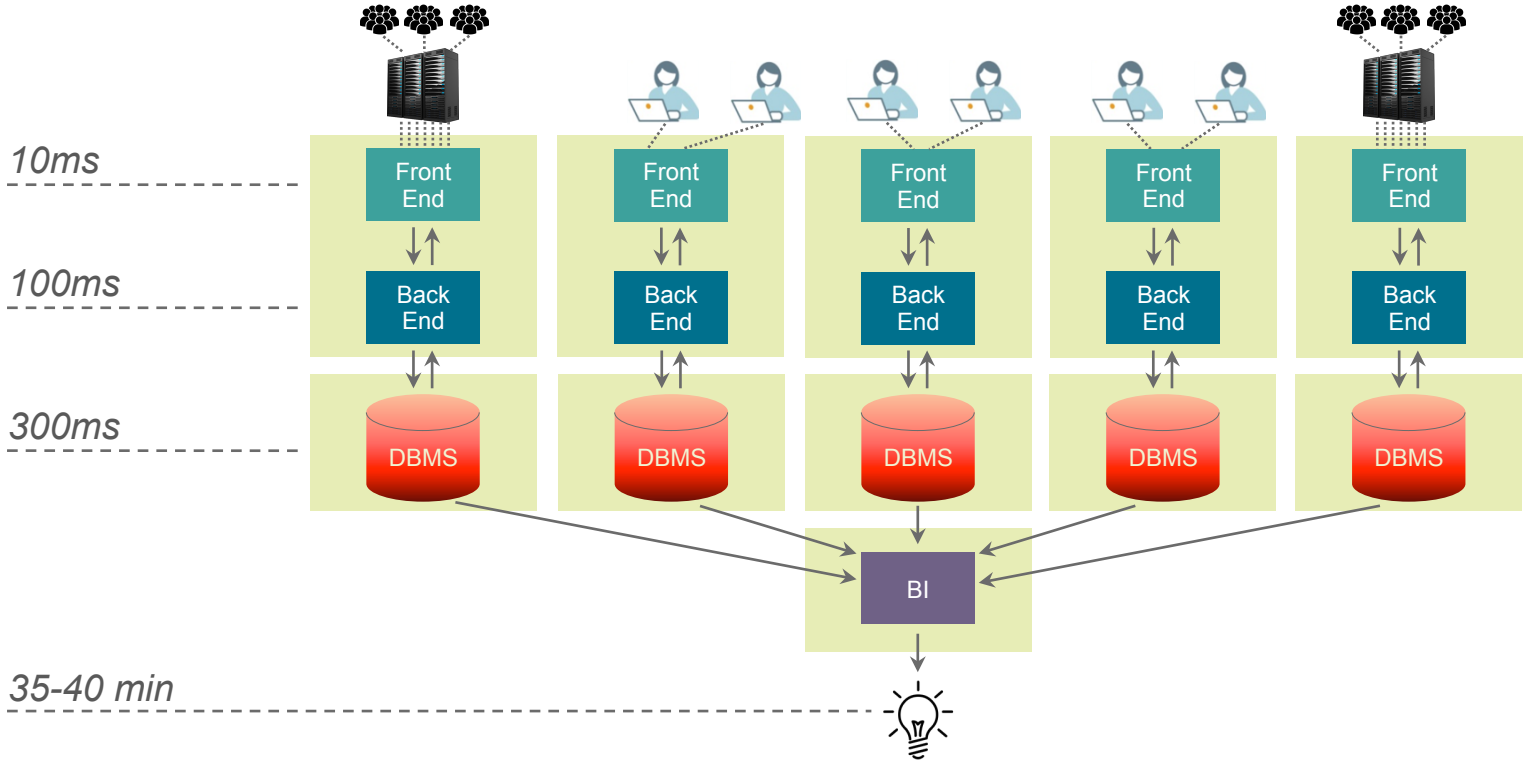
What if "split" didn't help this time?

First Issues

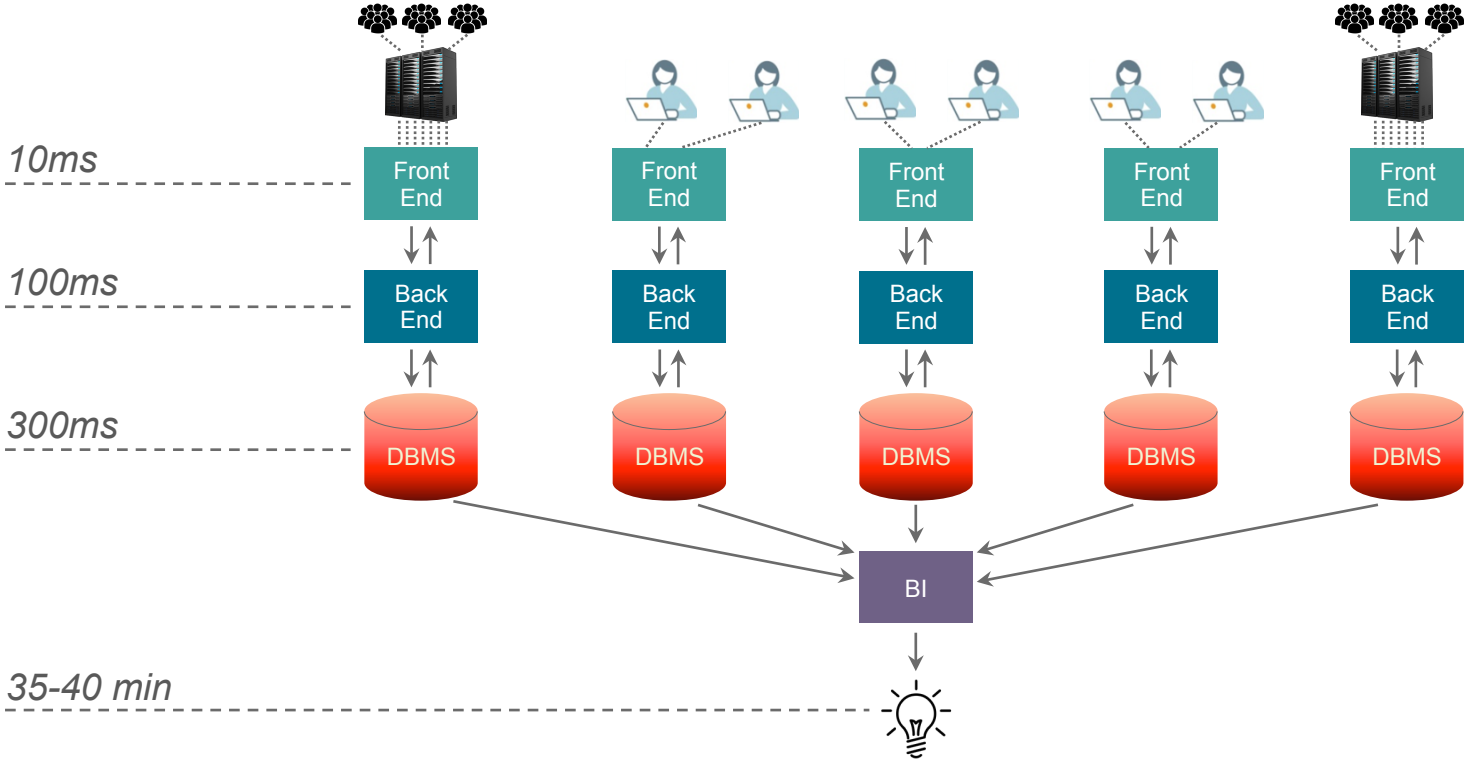


Split more! Eventually it will help…

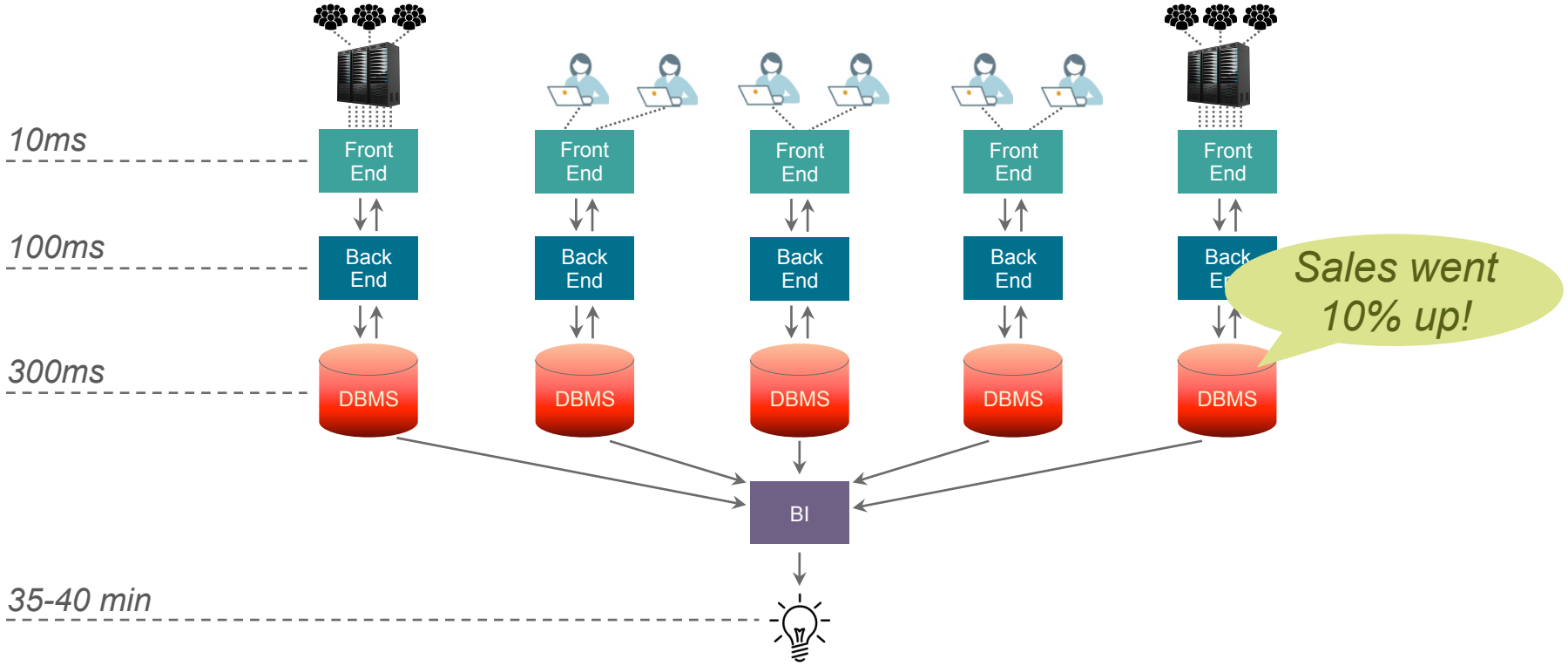
First Issues



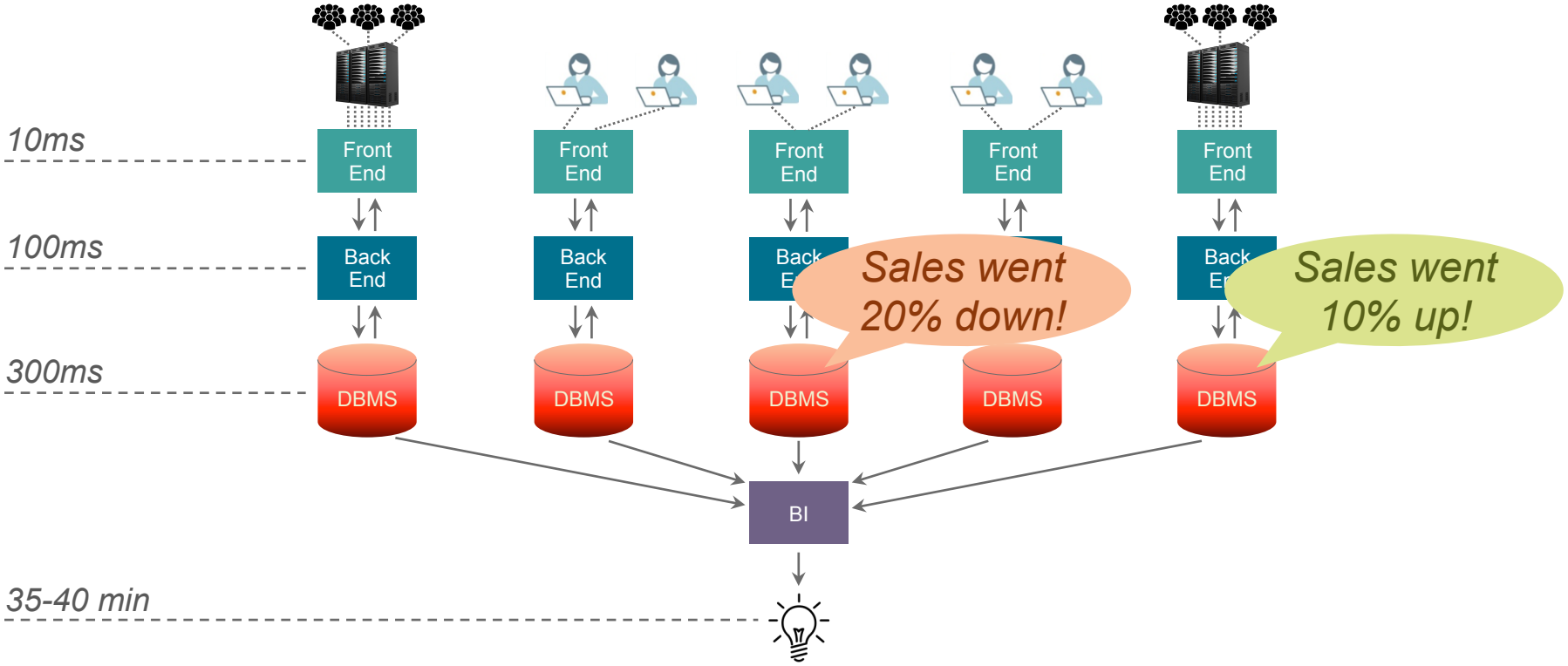
First Issues



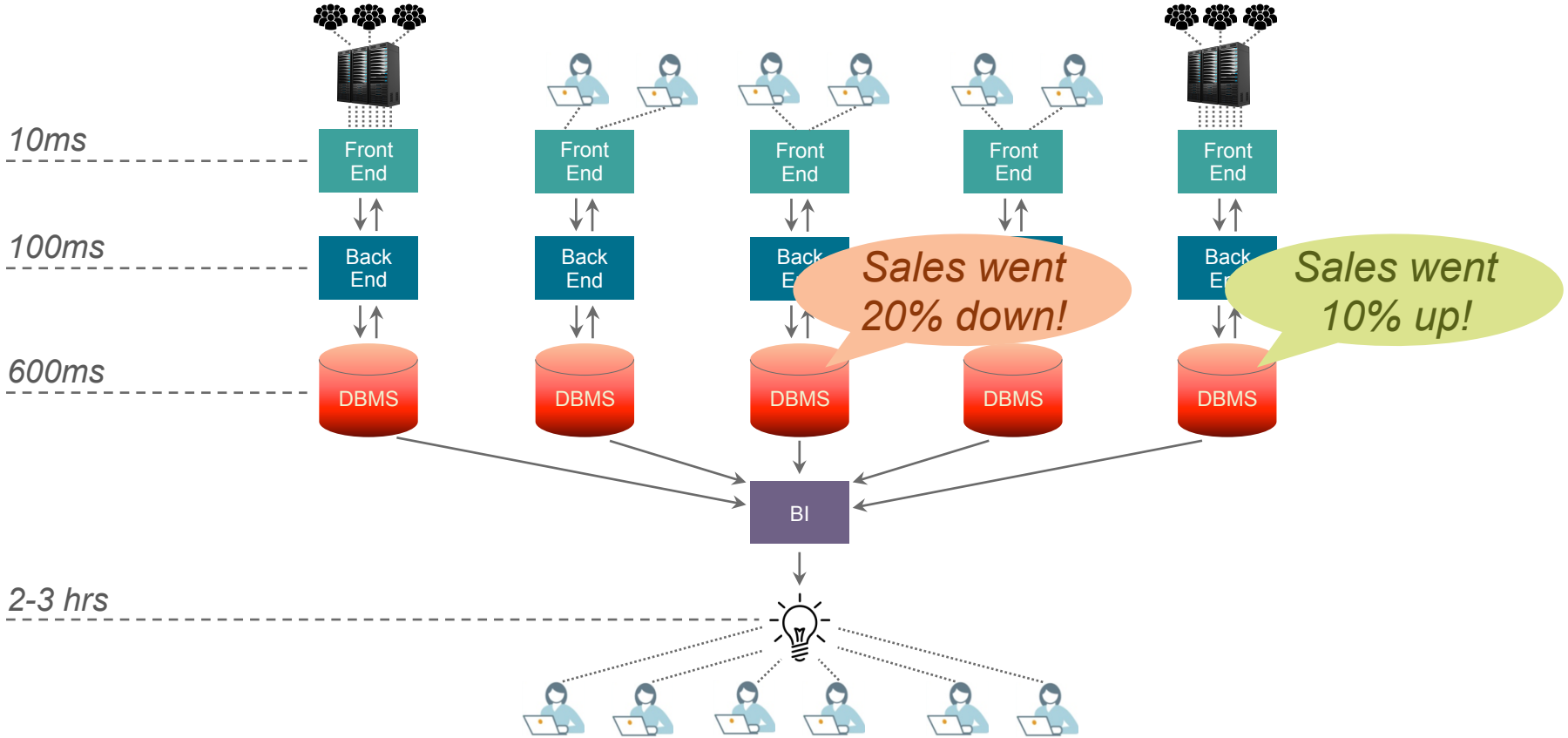
First Issues



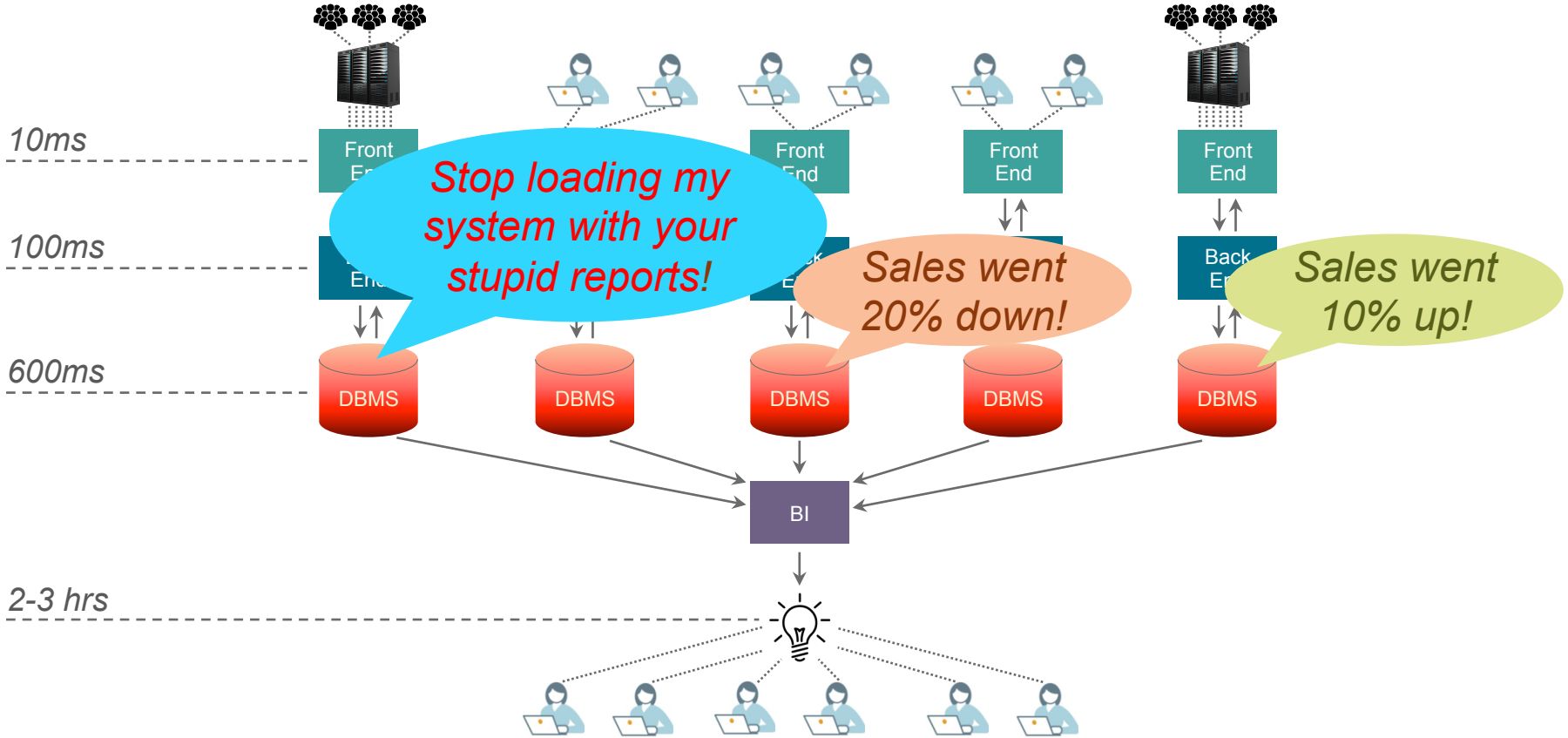
First Issues



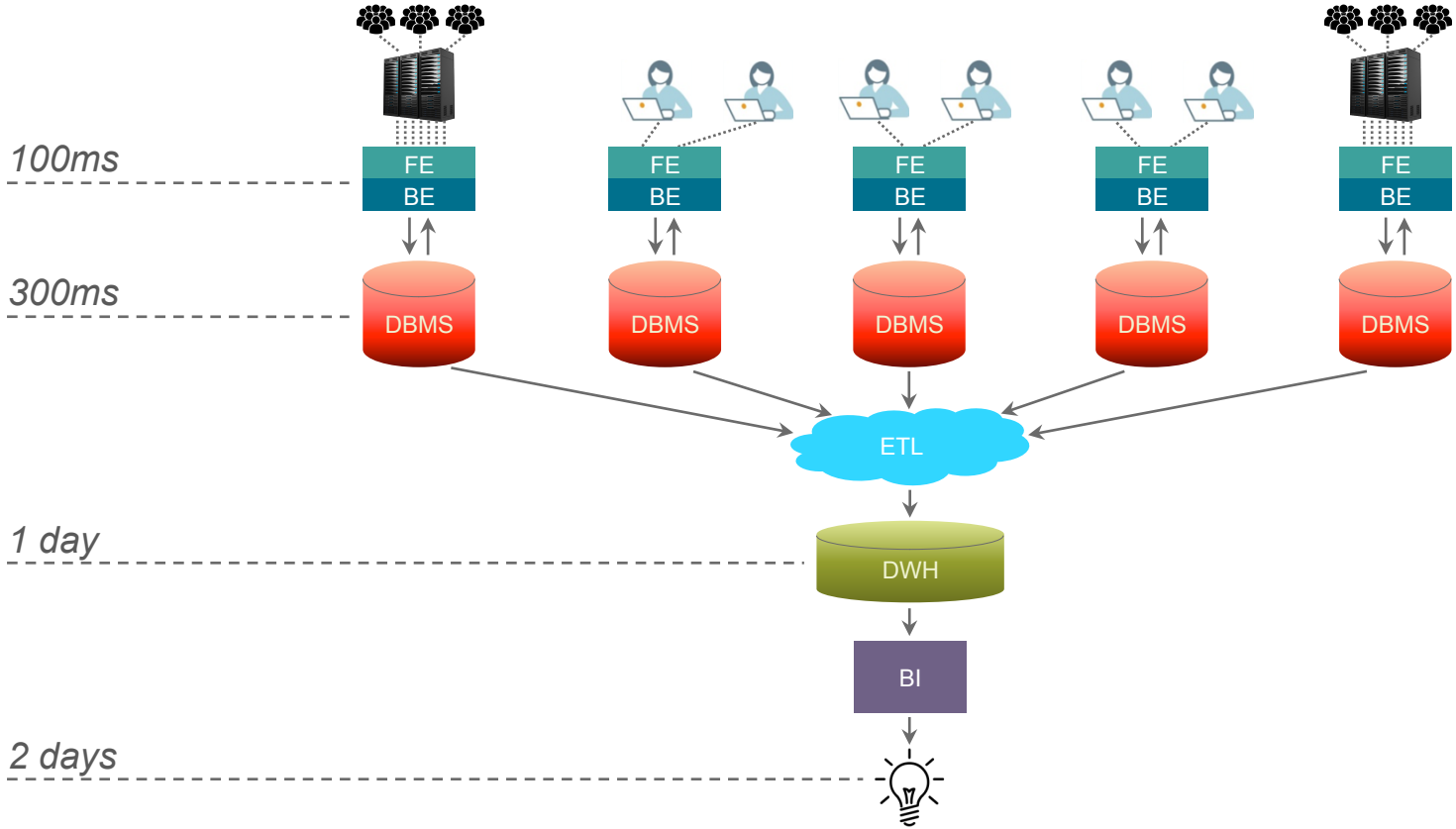
First Issues



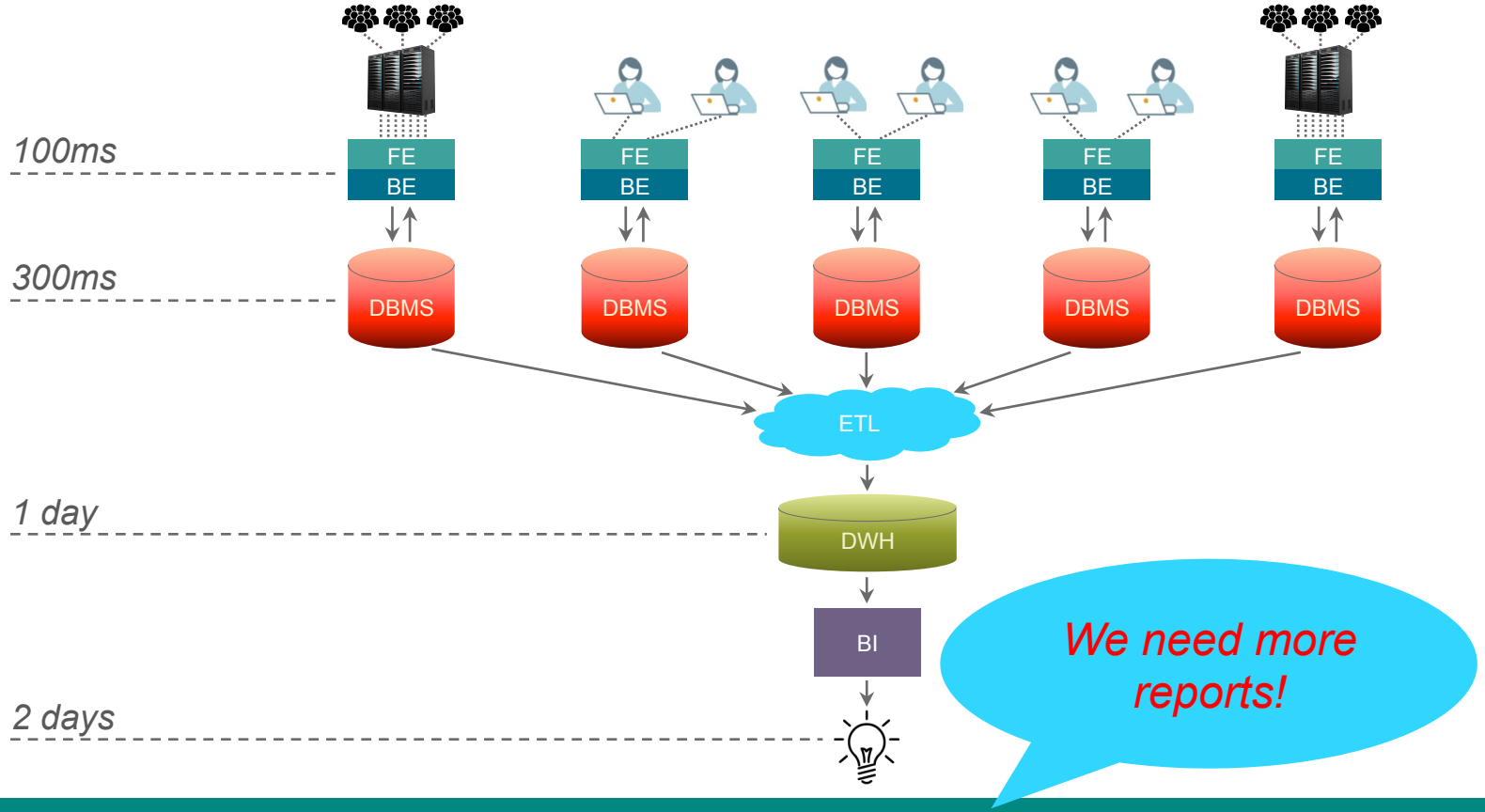
First Issues



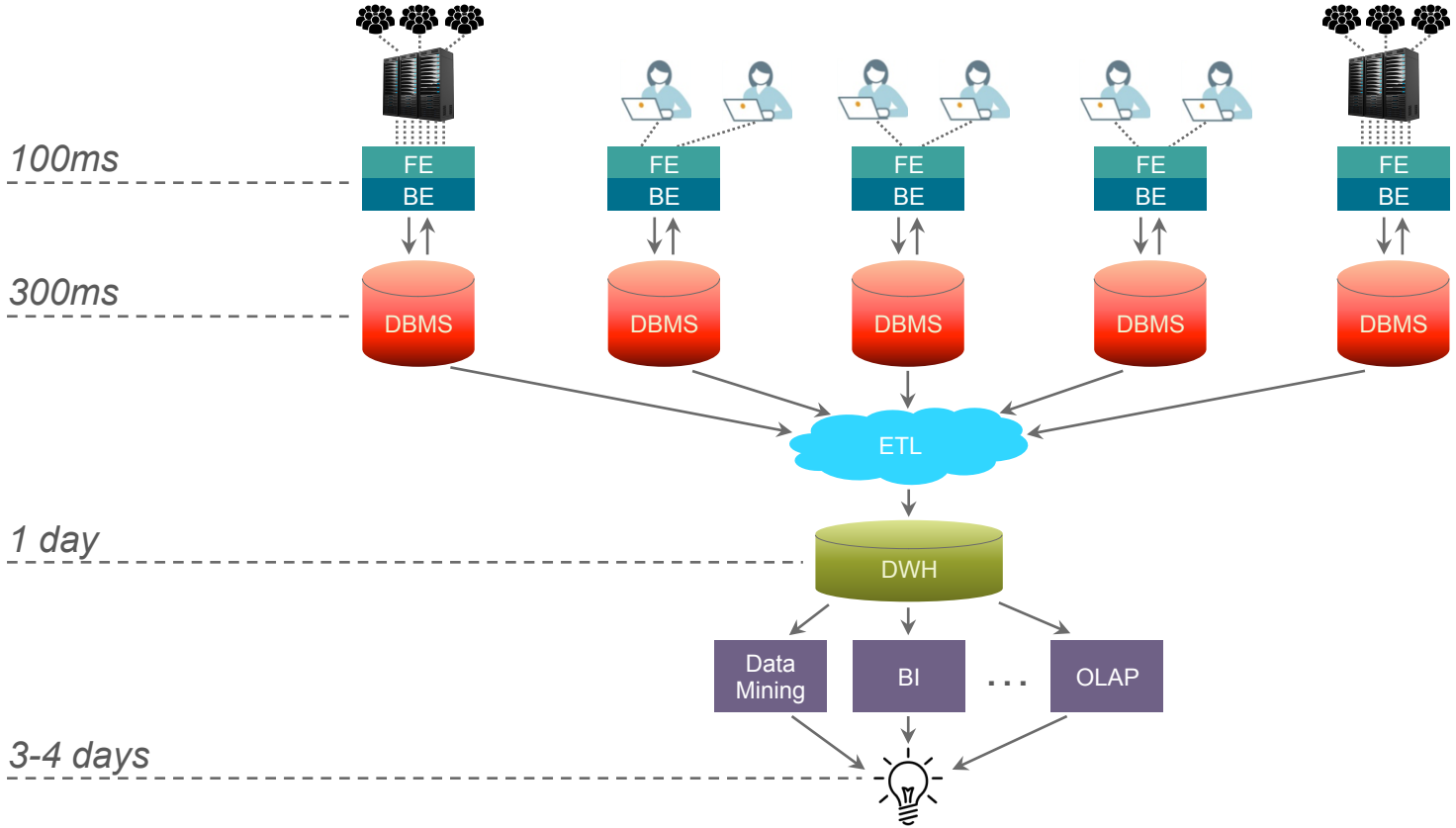
The Era of Data Warehouse



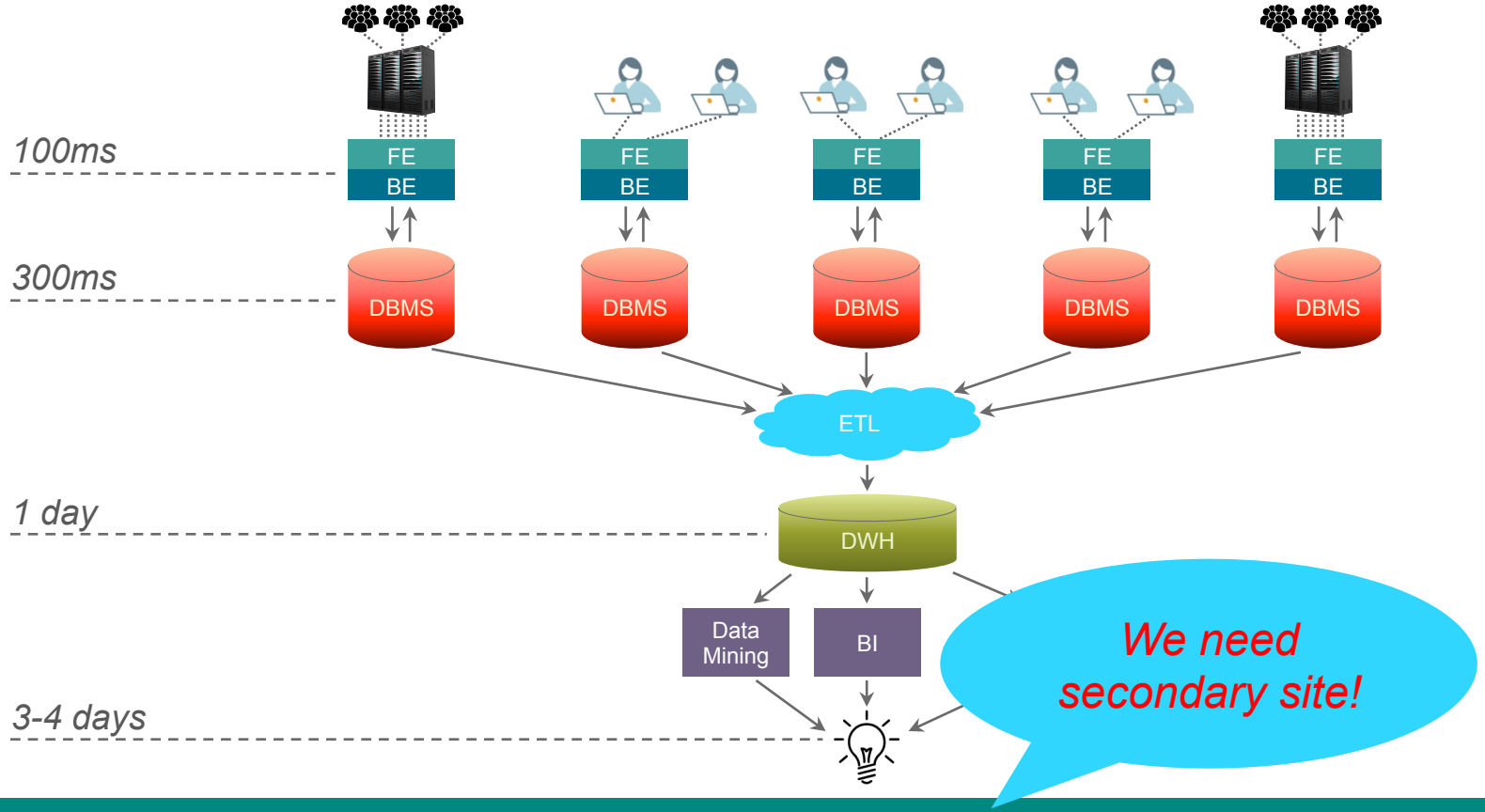
The Era of Data Warehouse



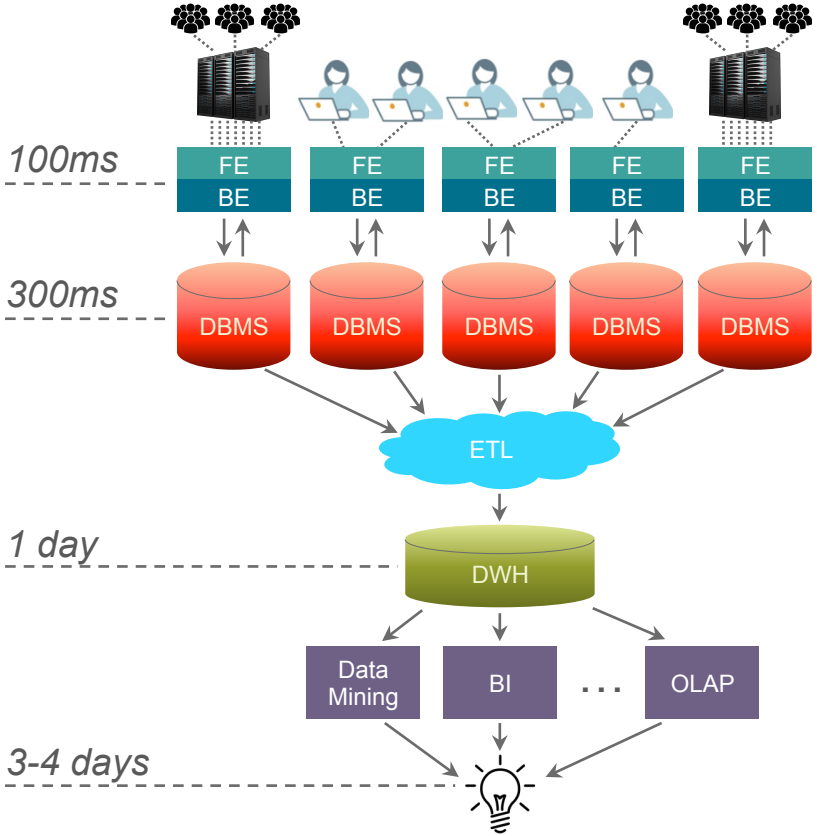
The Era of Data Warehouse



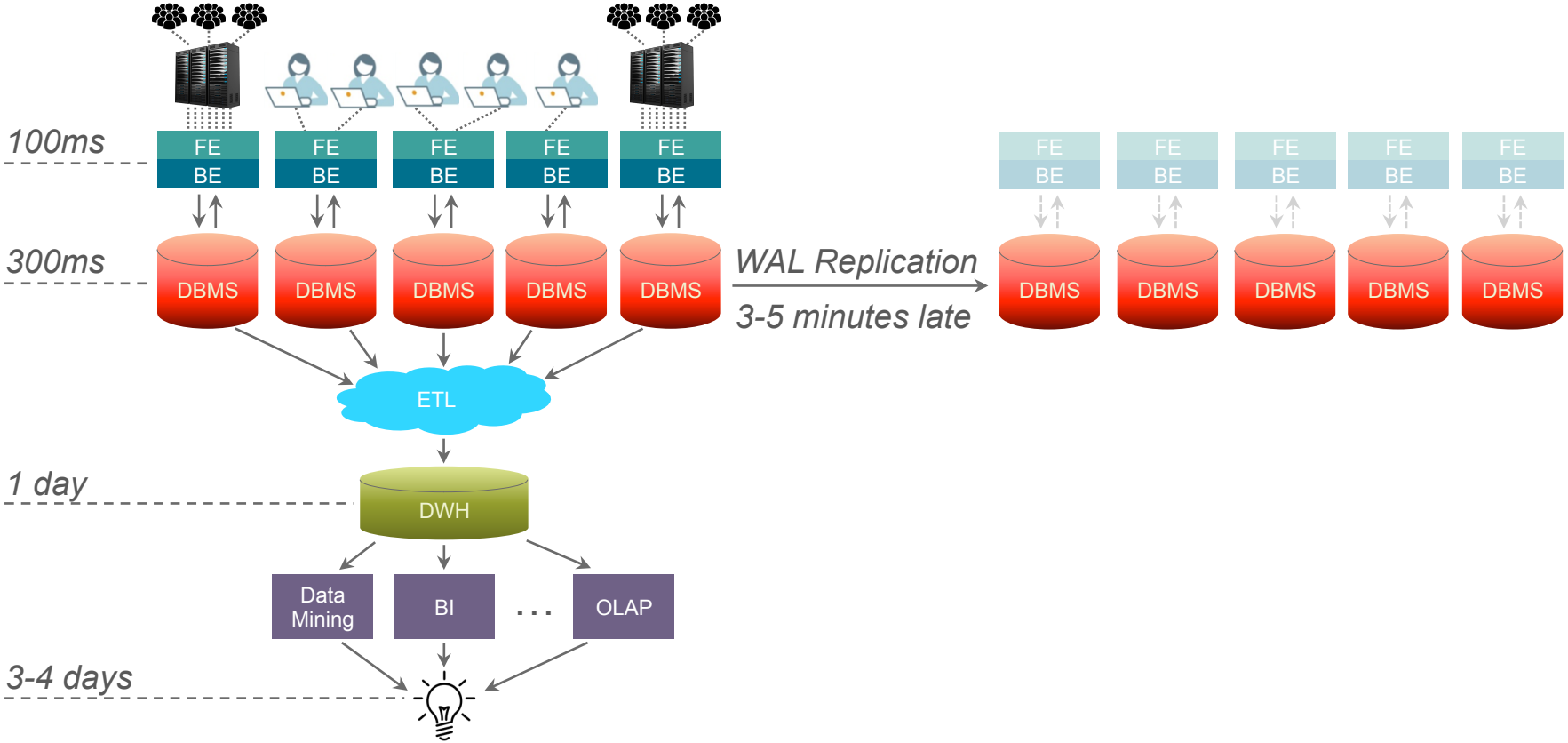
The Era of Data Warehouse



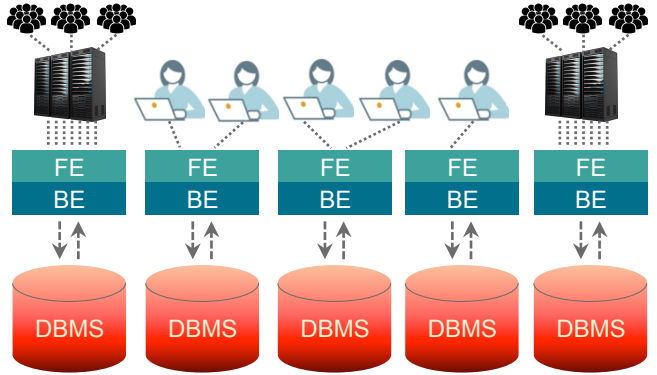
The Era of Data Warehouse



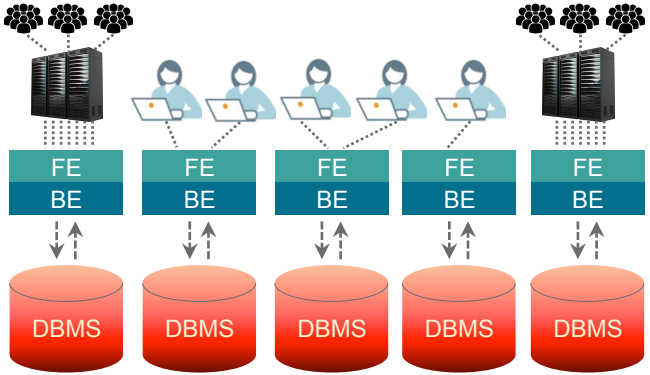
The Era of Data Warehouse



The Era of Data Warehouse

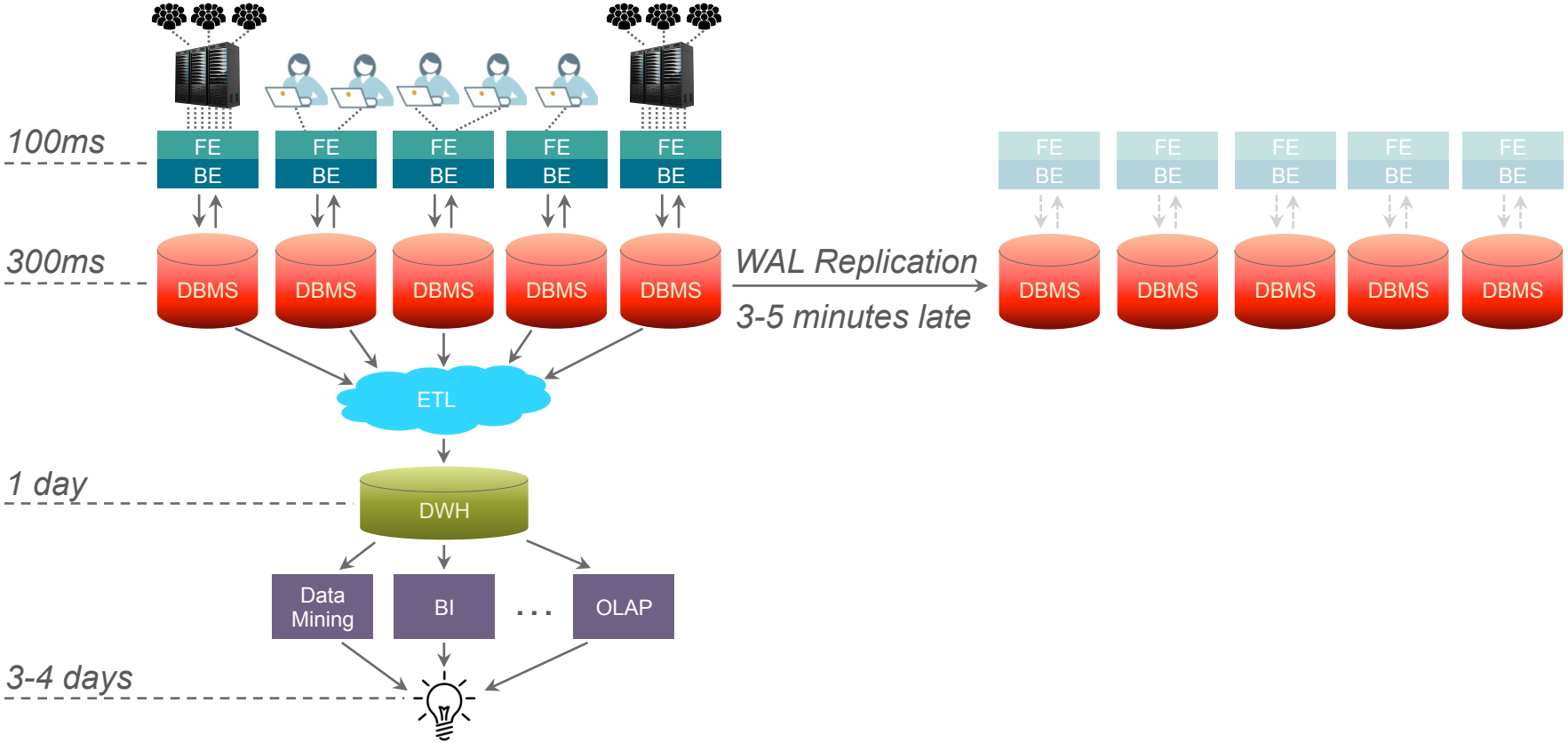


The Era of Data Warehouse

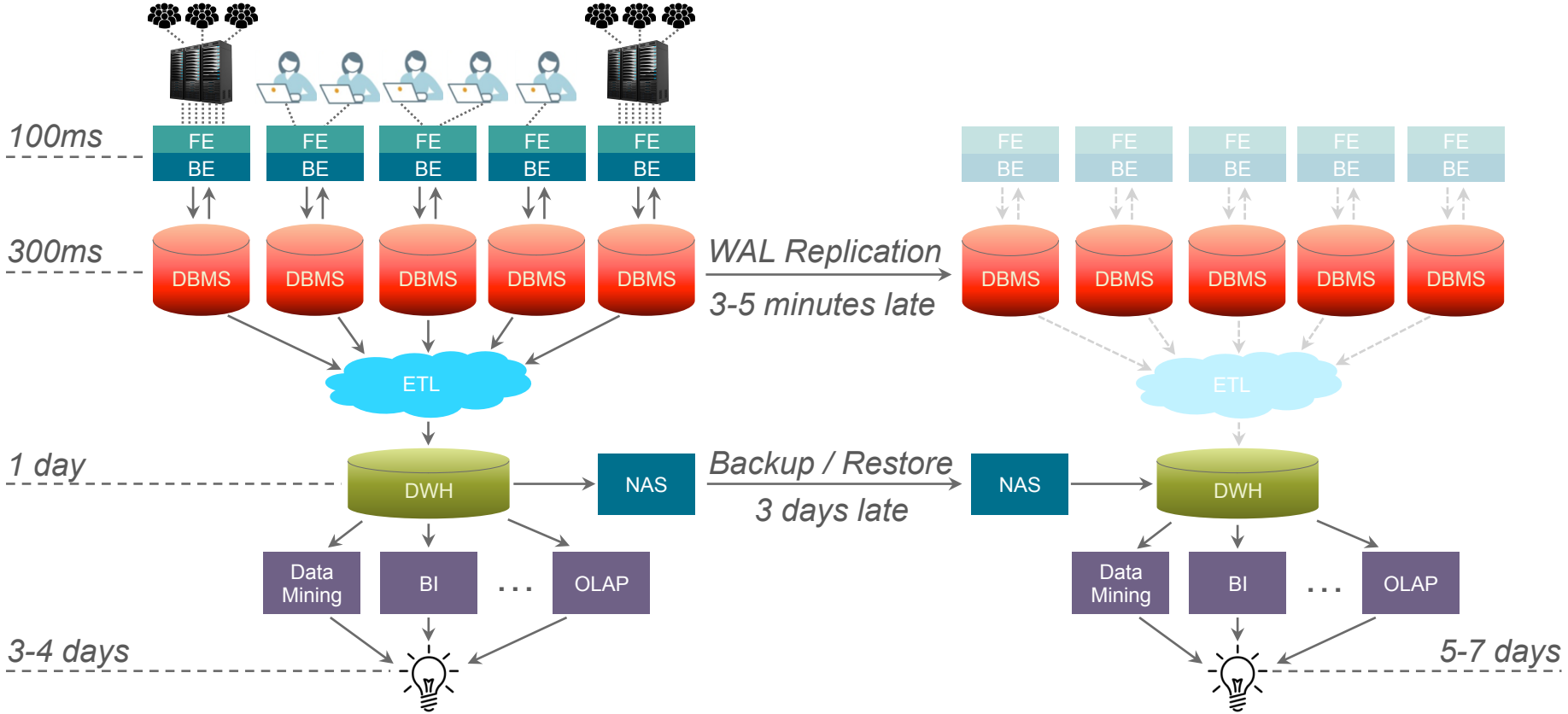


Where is our DWH? We need this data now!

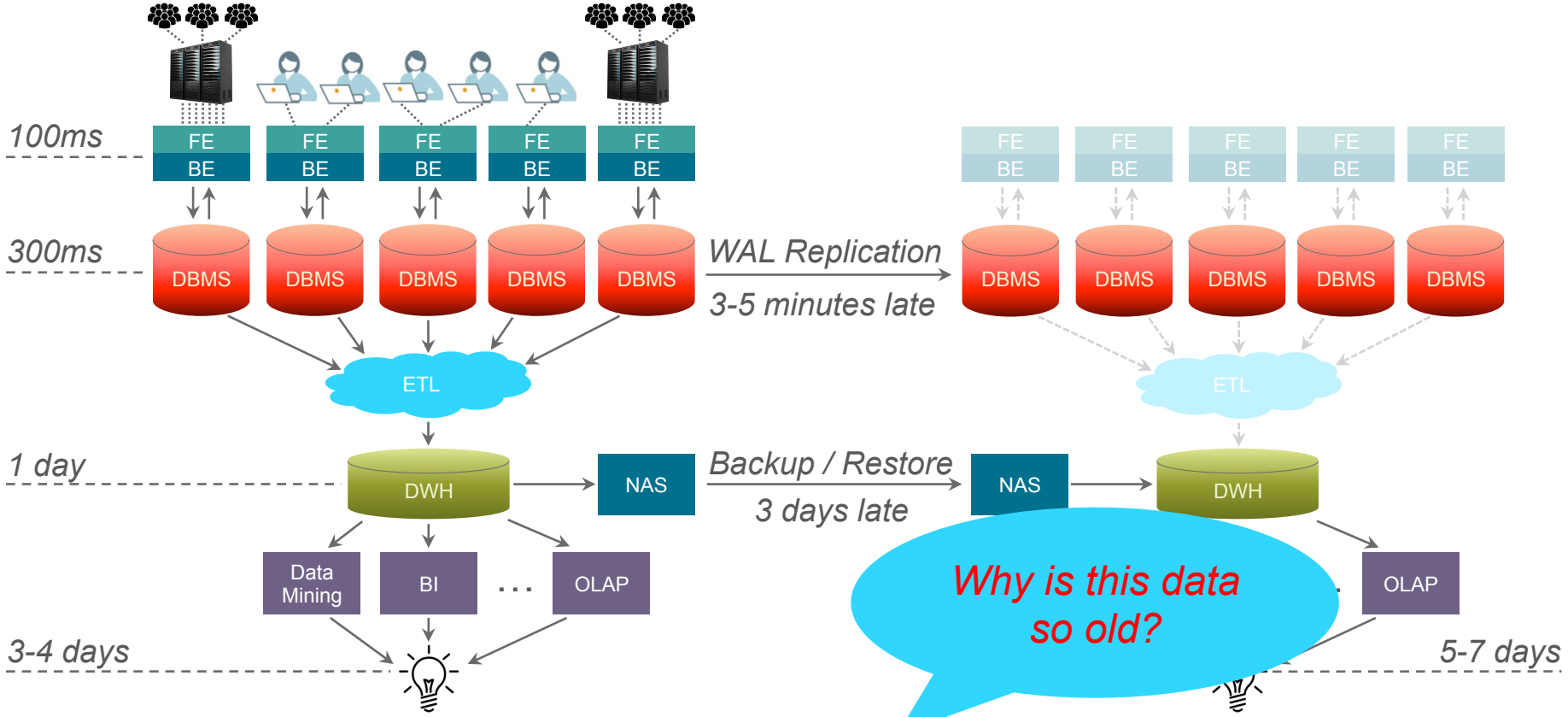
The Era of Data Warehouse



The Era of Data Warehouse

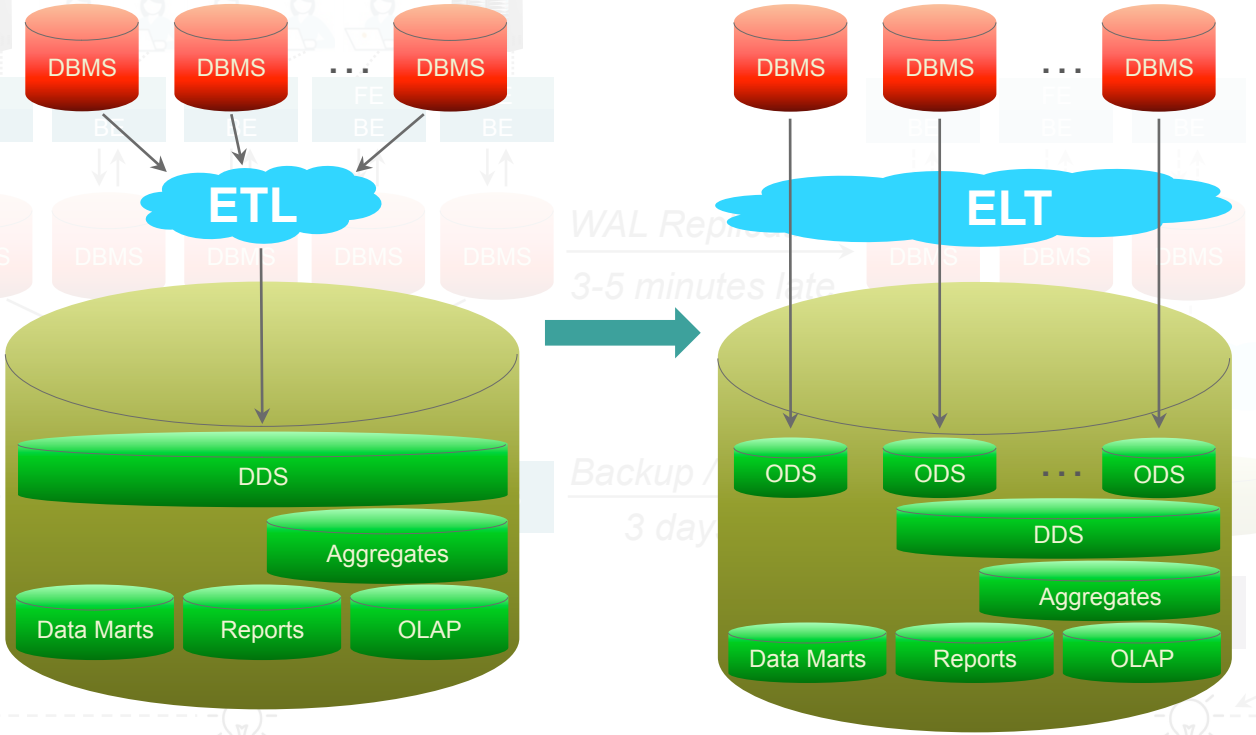


The Era of Data Warehouse

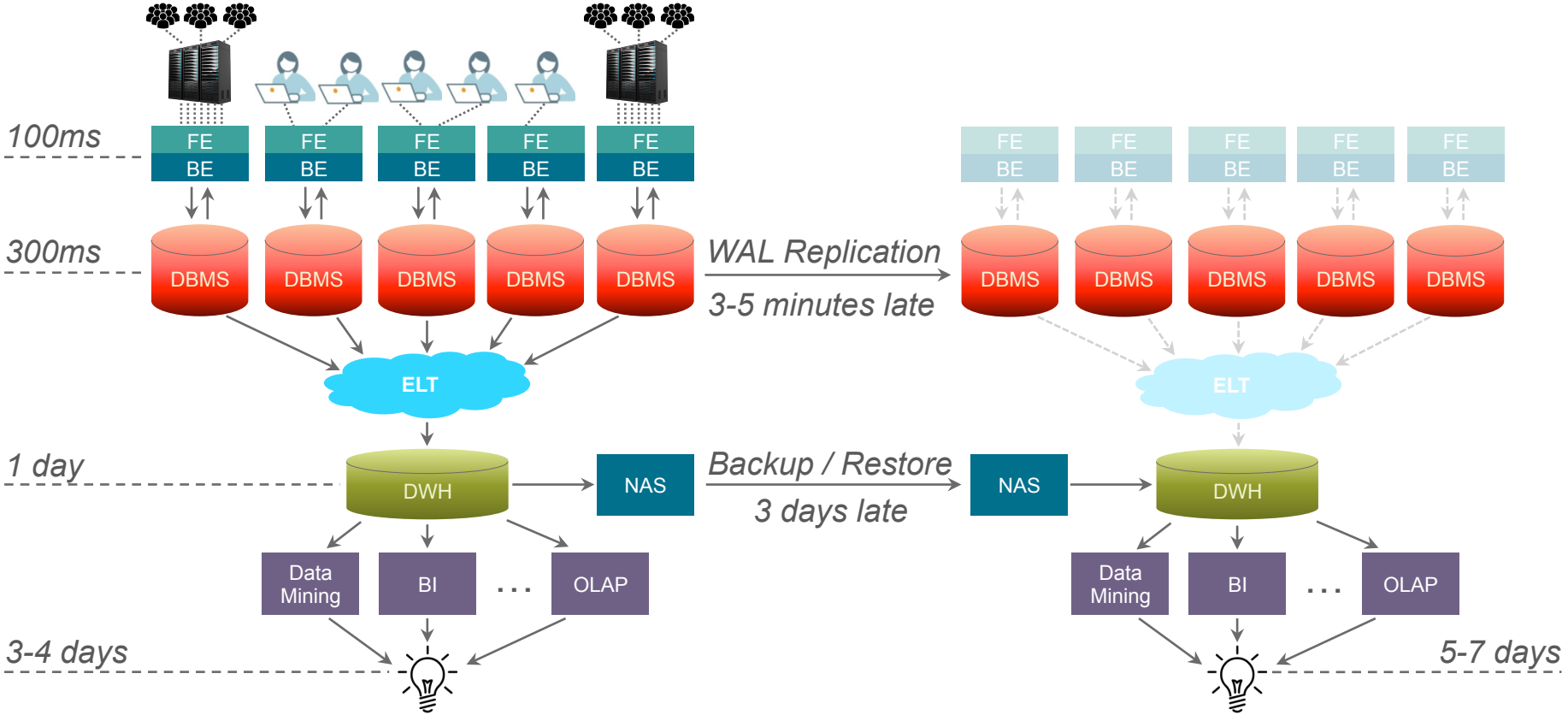


The Era of Data Warehouse

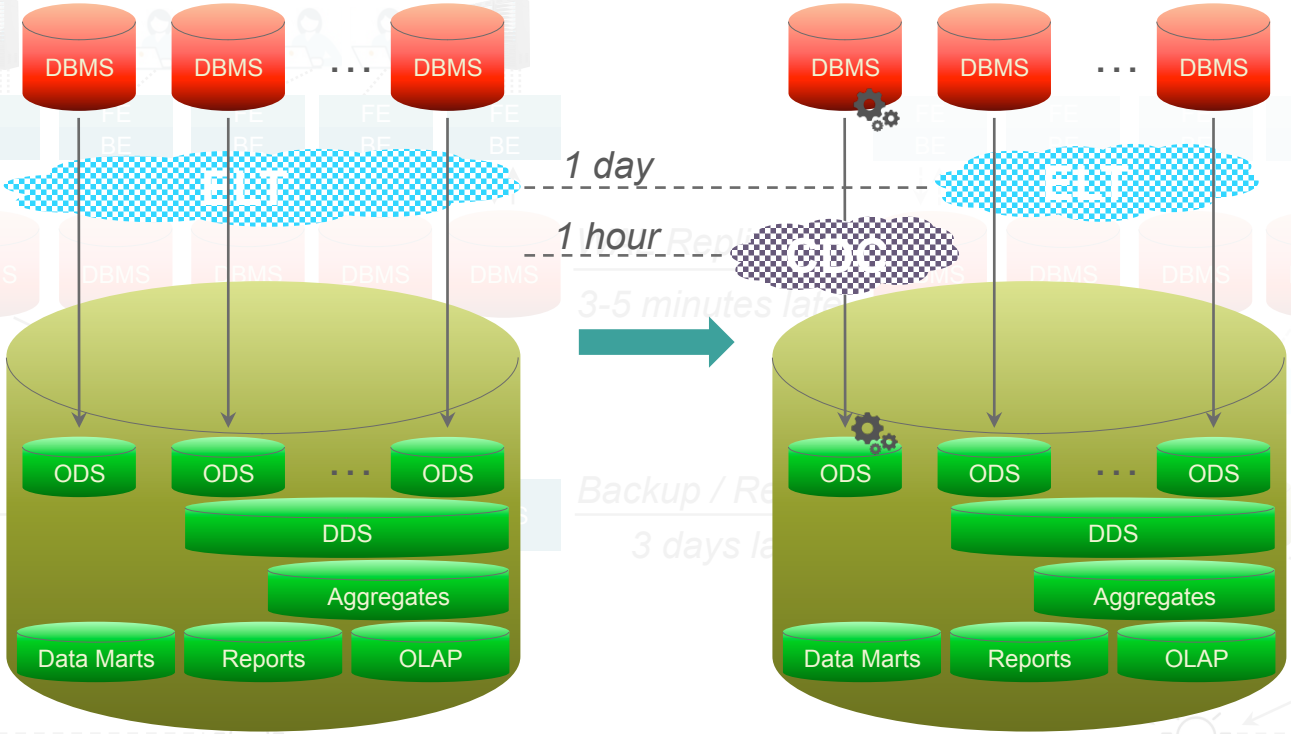
Advanced Architecture – ELT



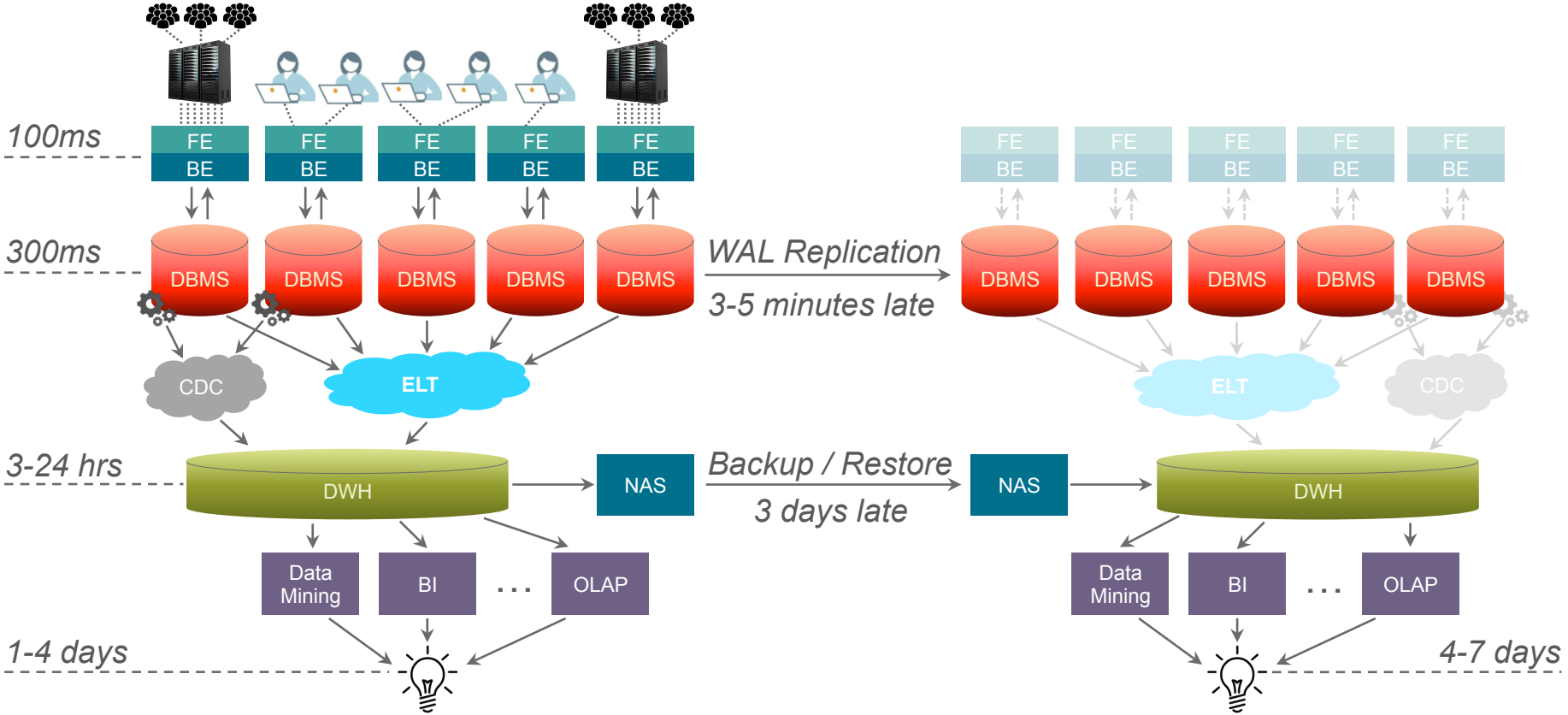
Advanced Architecture – ELT



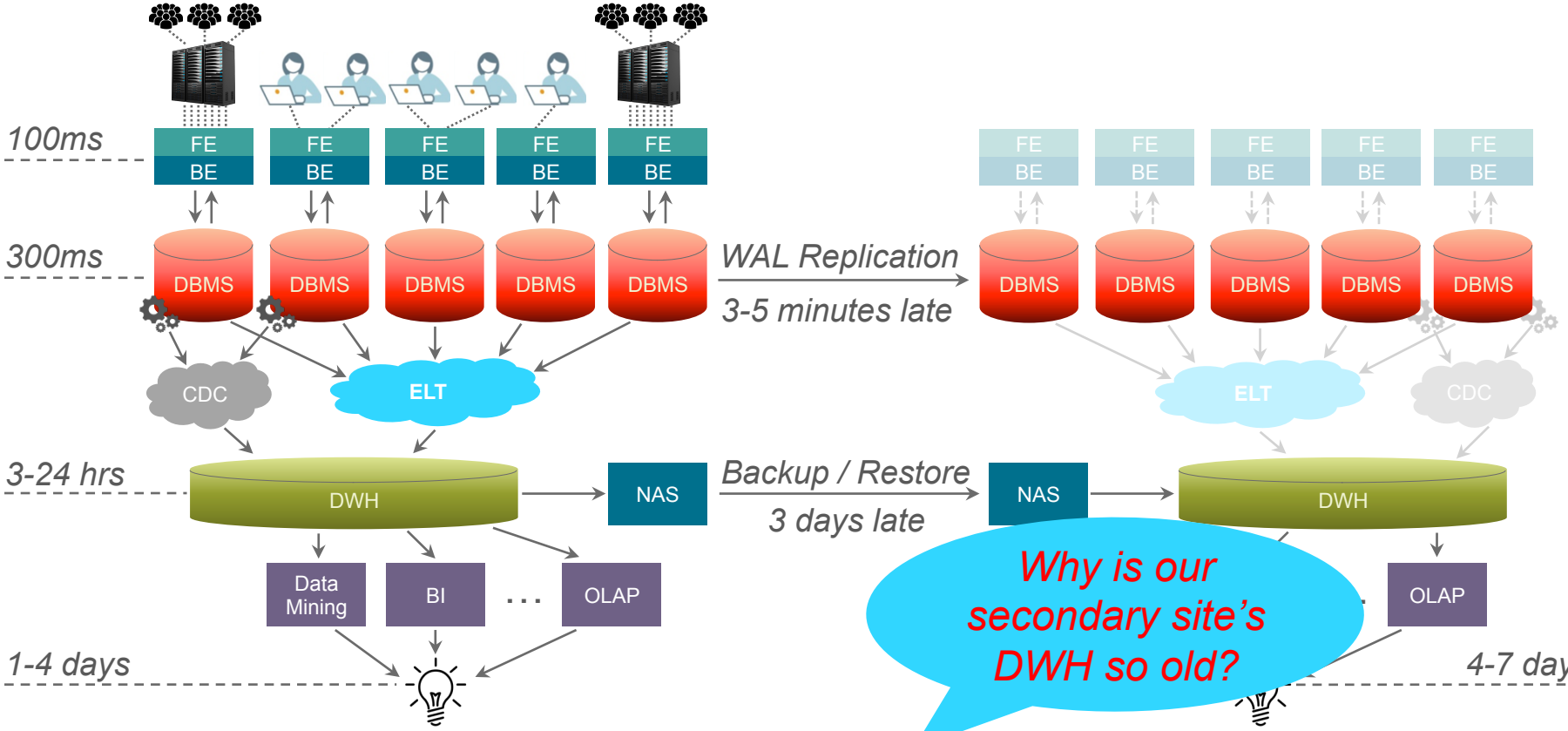
Advanced Architecture – CDC



Advanced Architecture – CDC



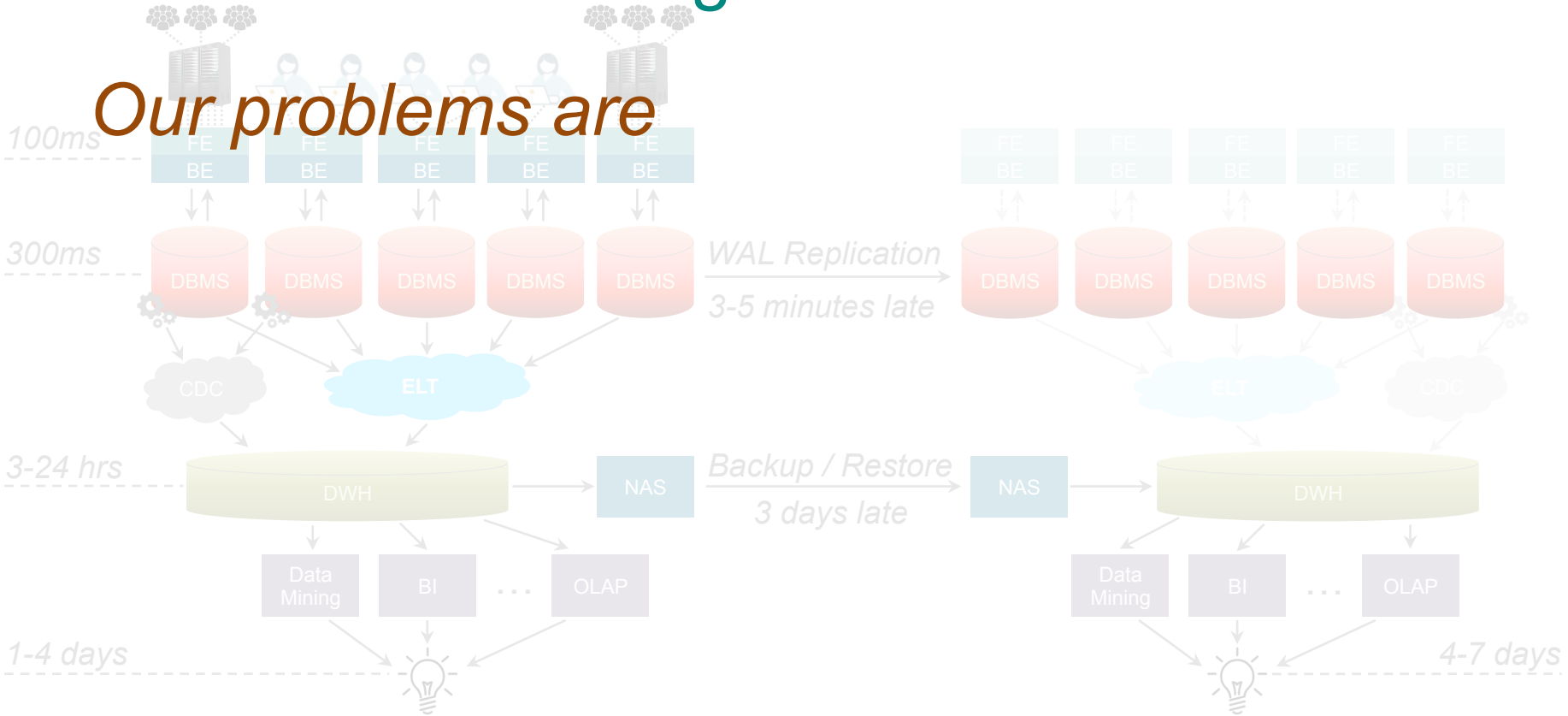
Advanced Architecture – CDC



Moving Forward

Moving Forward

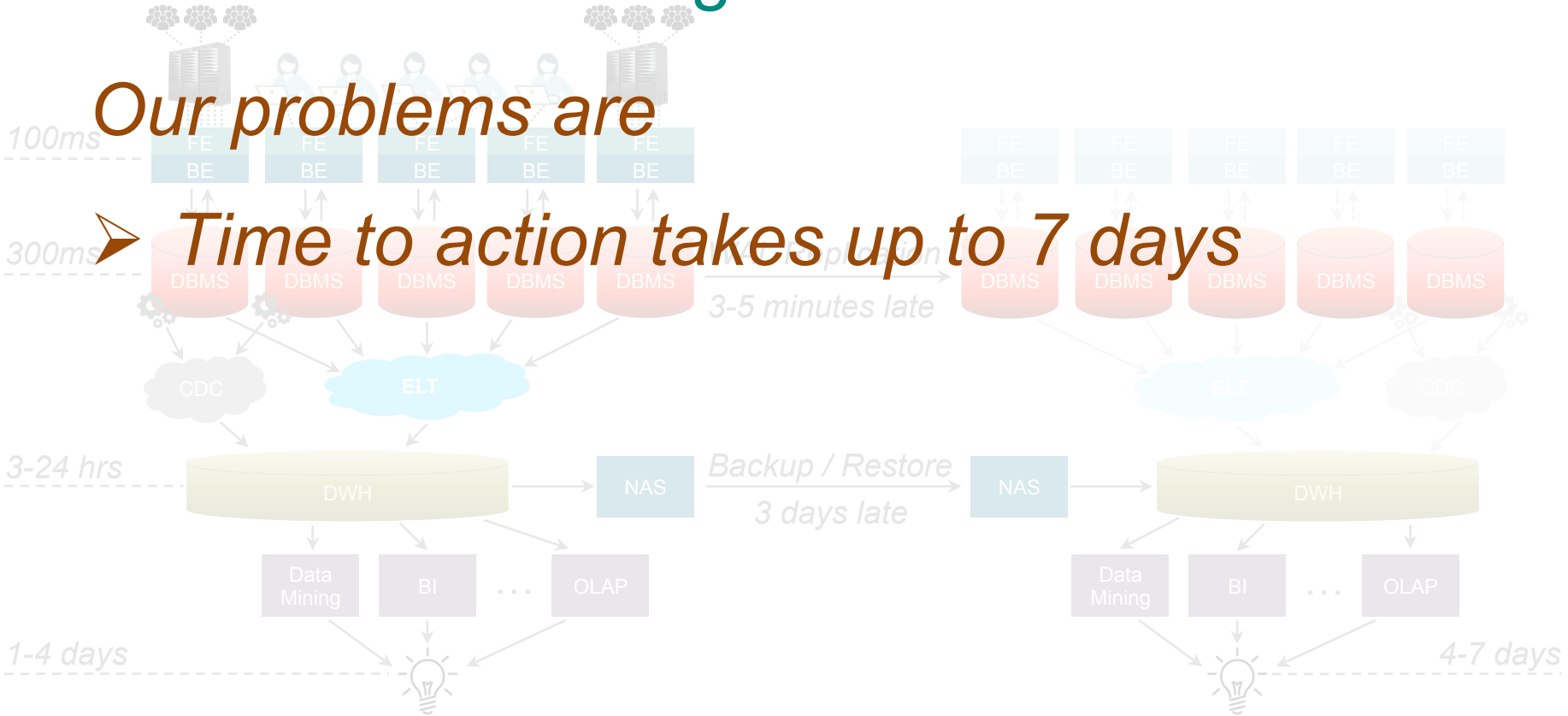
Our problems are



Moving Forward

Our problems are

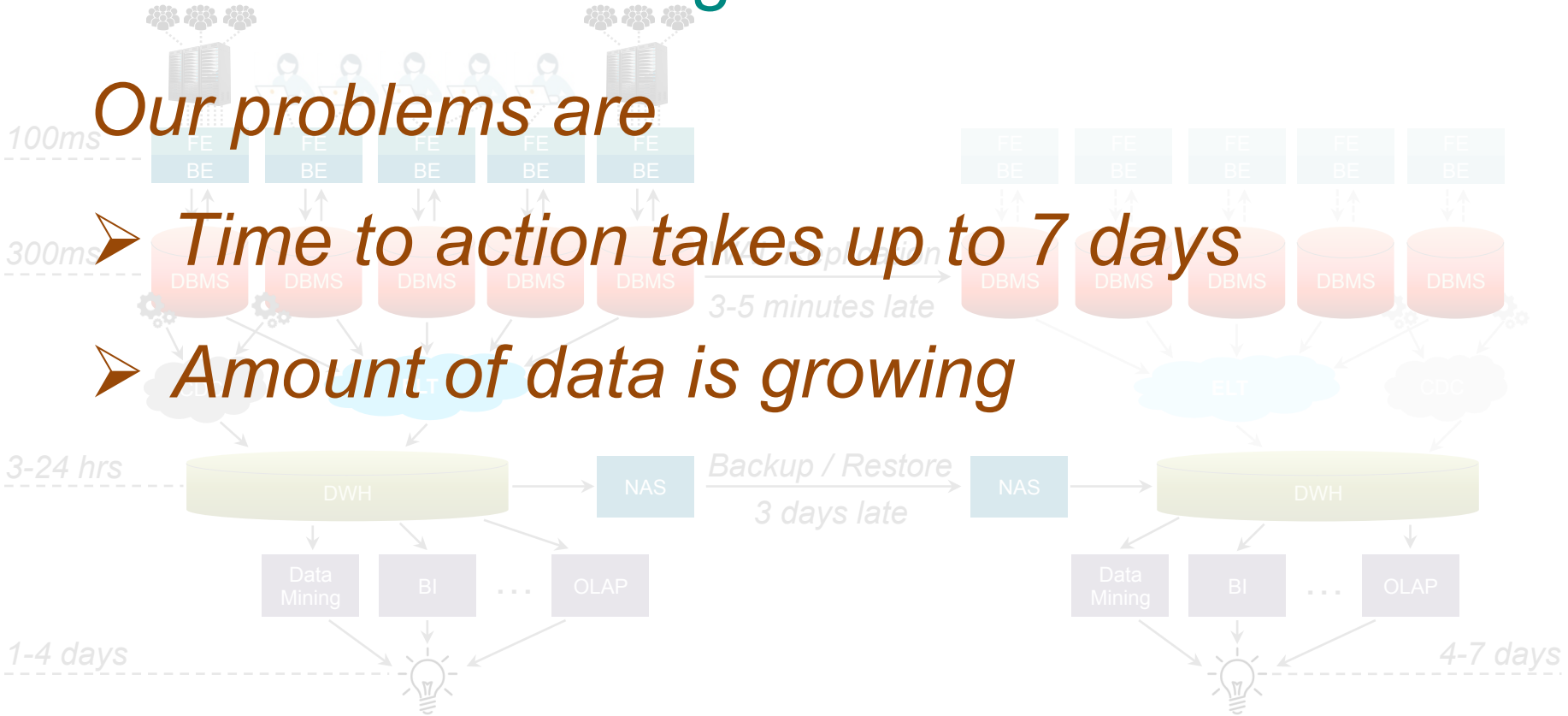
- *Time to action takes up to 7 days*



Moving Forward

Our problems are

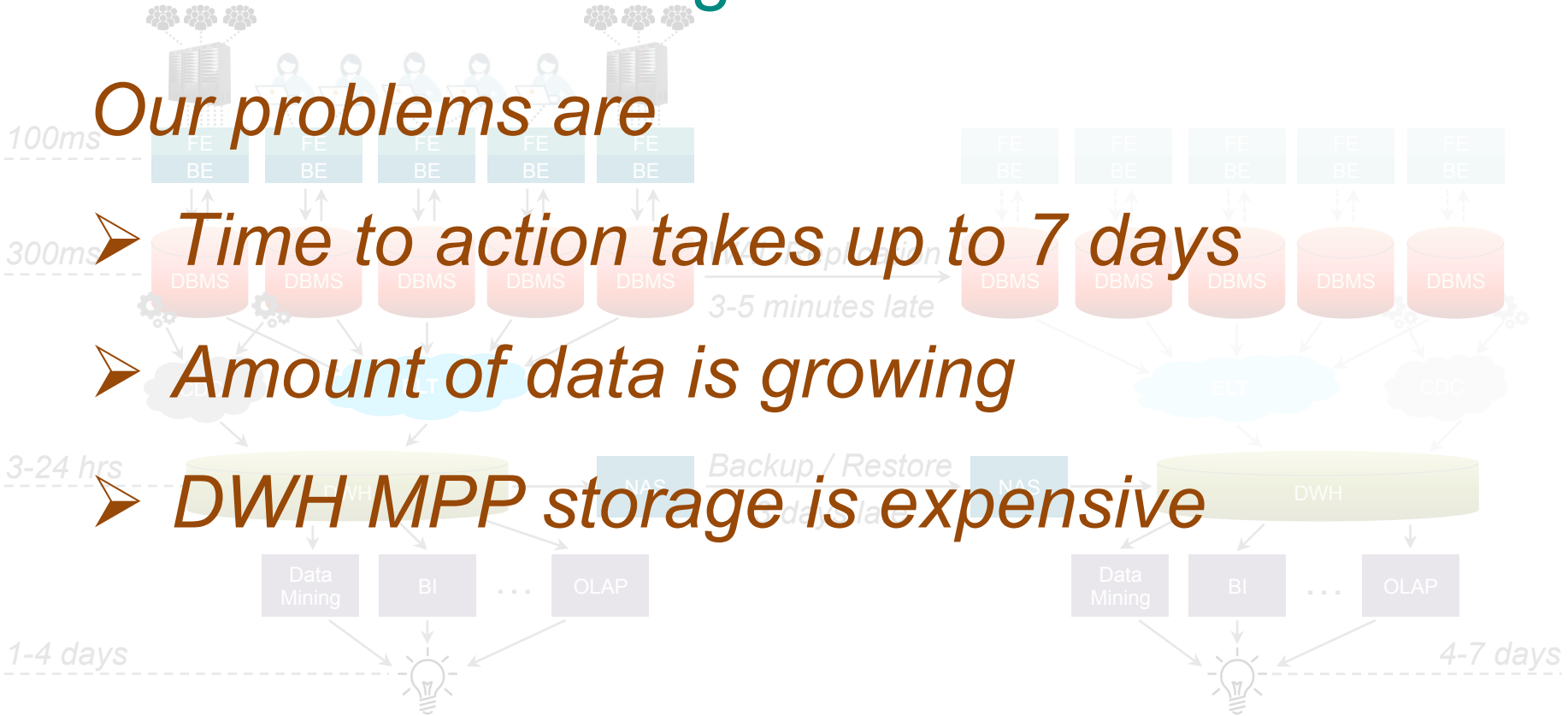
- *Time to action takes up to 7 days*
- *Amount of data is growing*



Moving Forward

Our problems are

- *Time to action takes up to 7 days*
- *Amount of data is growing*
- *DWH MPP storage is expensive*

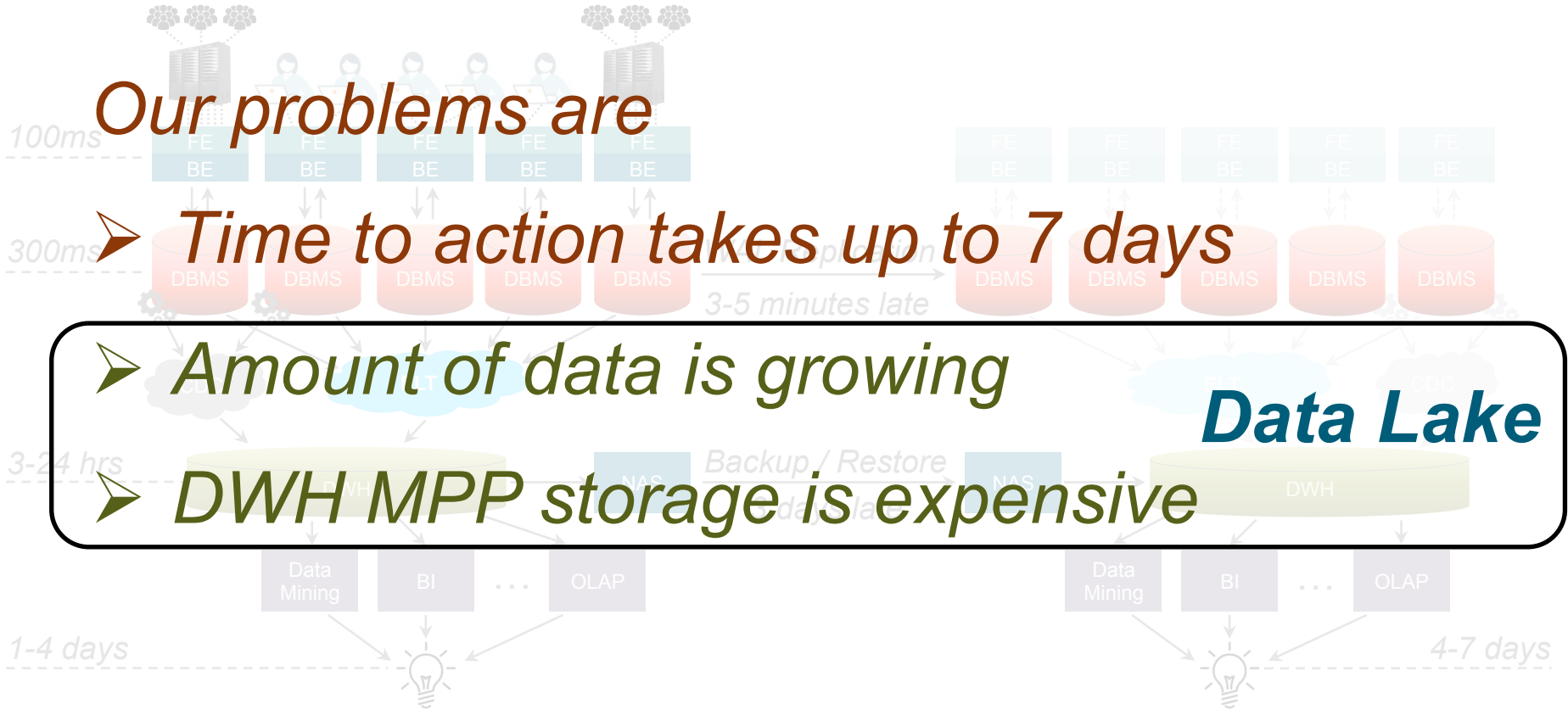


Modern Architectures

Our problems are

- *Time to action takes up to 7 days*
- *Amount of data is growing*
- *DWH MPP storage is expensive*

Data Lake



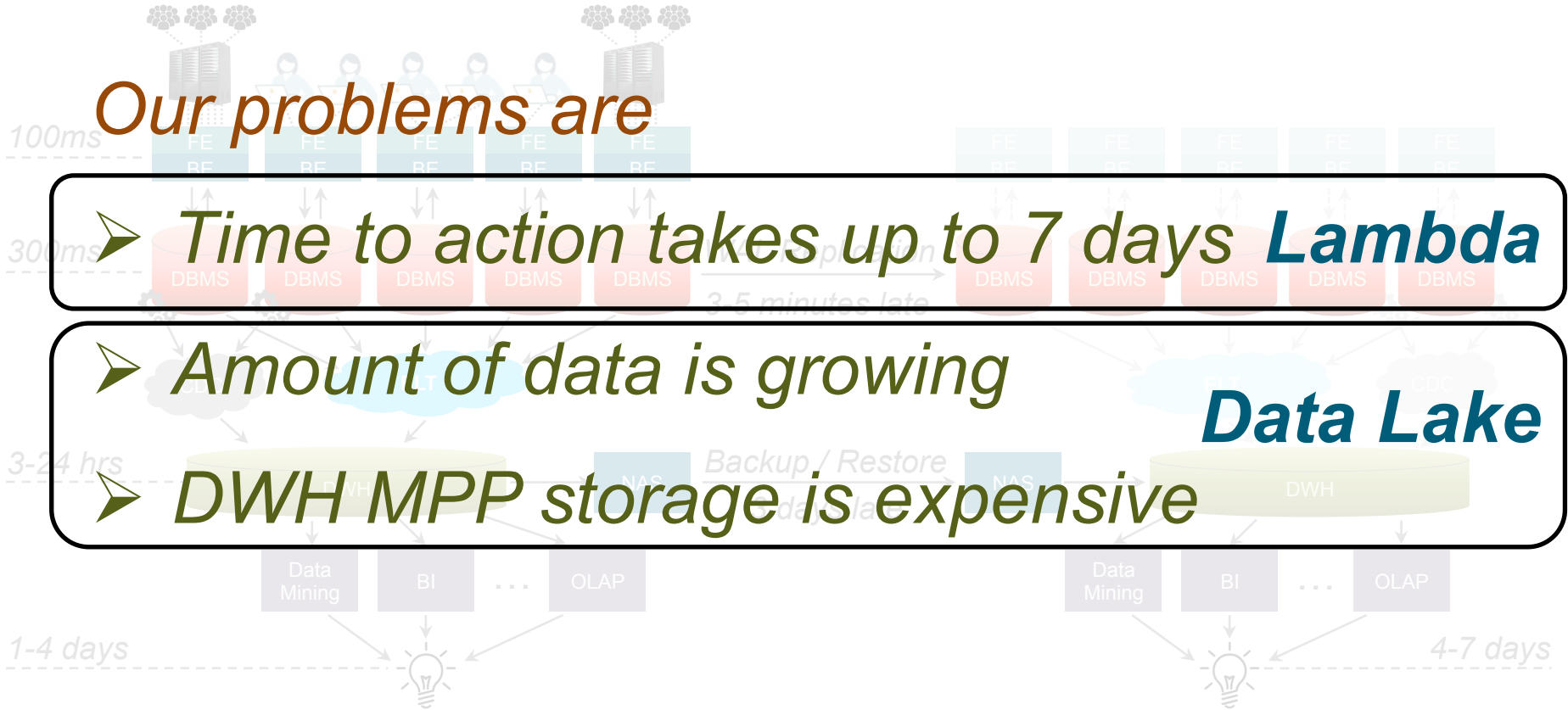
Modern Architectures

Our problems are

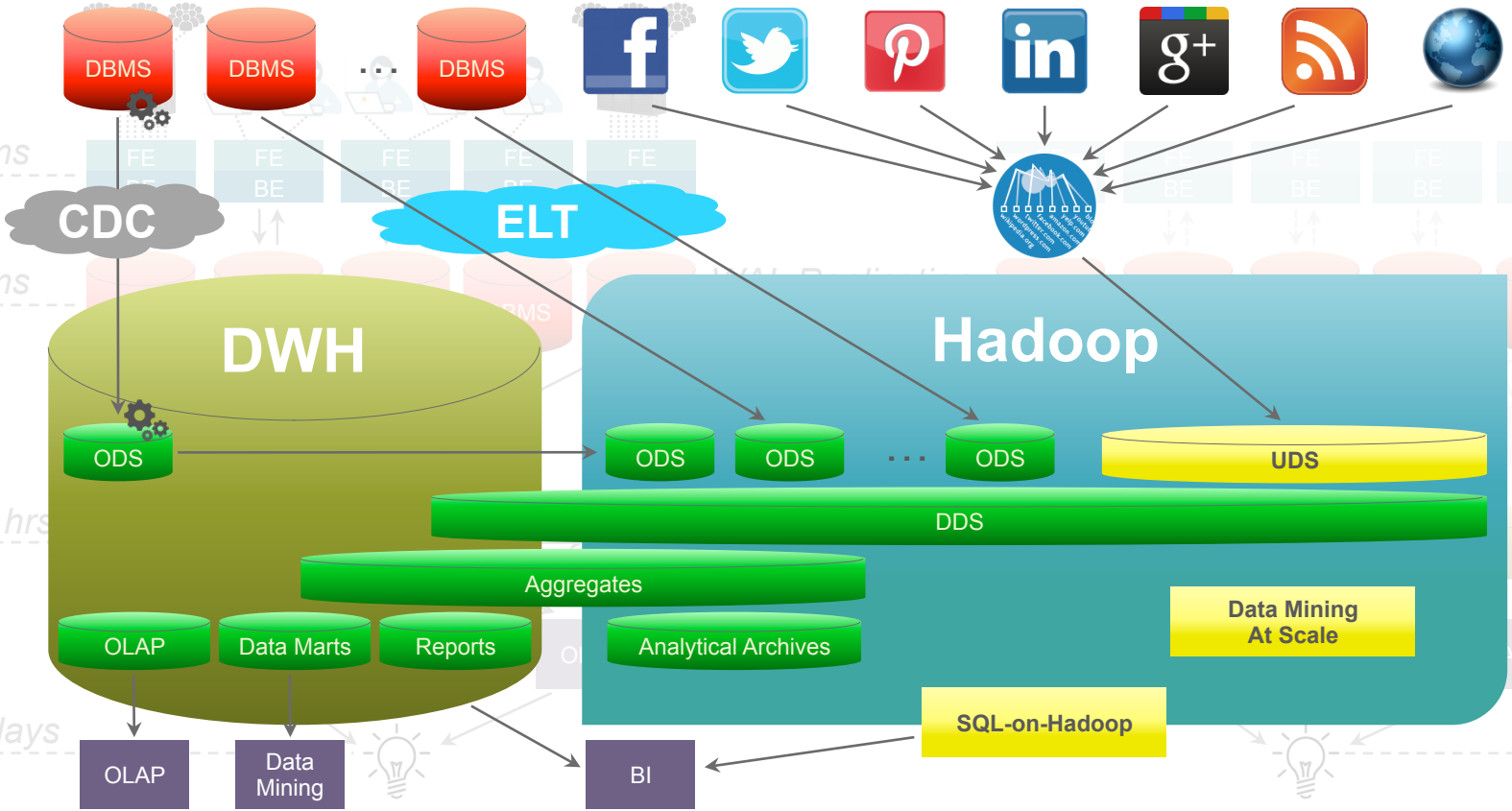
- *Time to action takes up to 7 days* **Lambda**

- *Amount of data is growing*
- *DWH MPP storage is expensive*

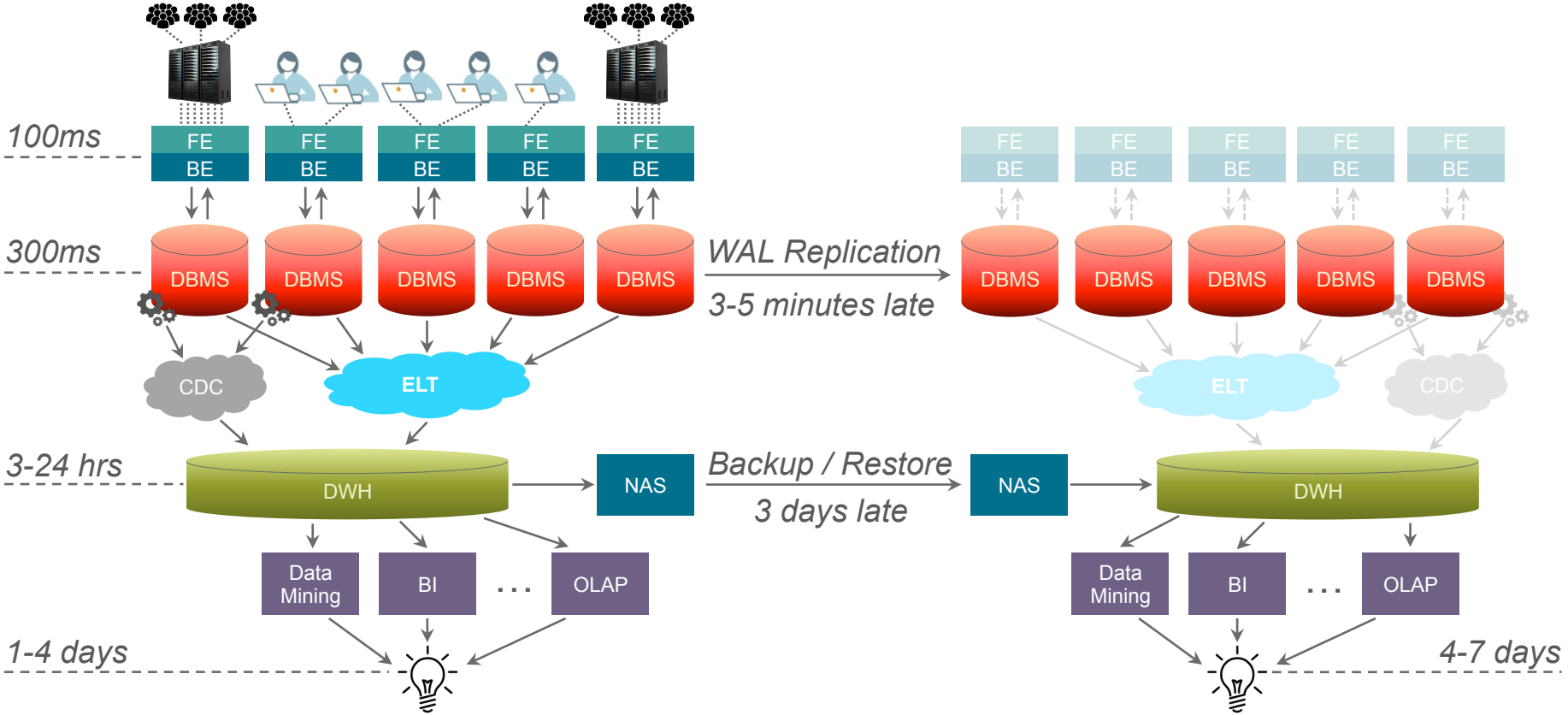
Data Lake



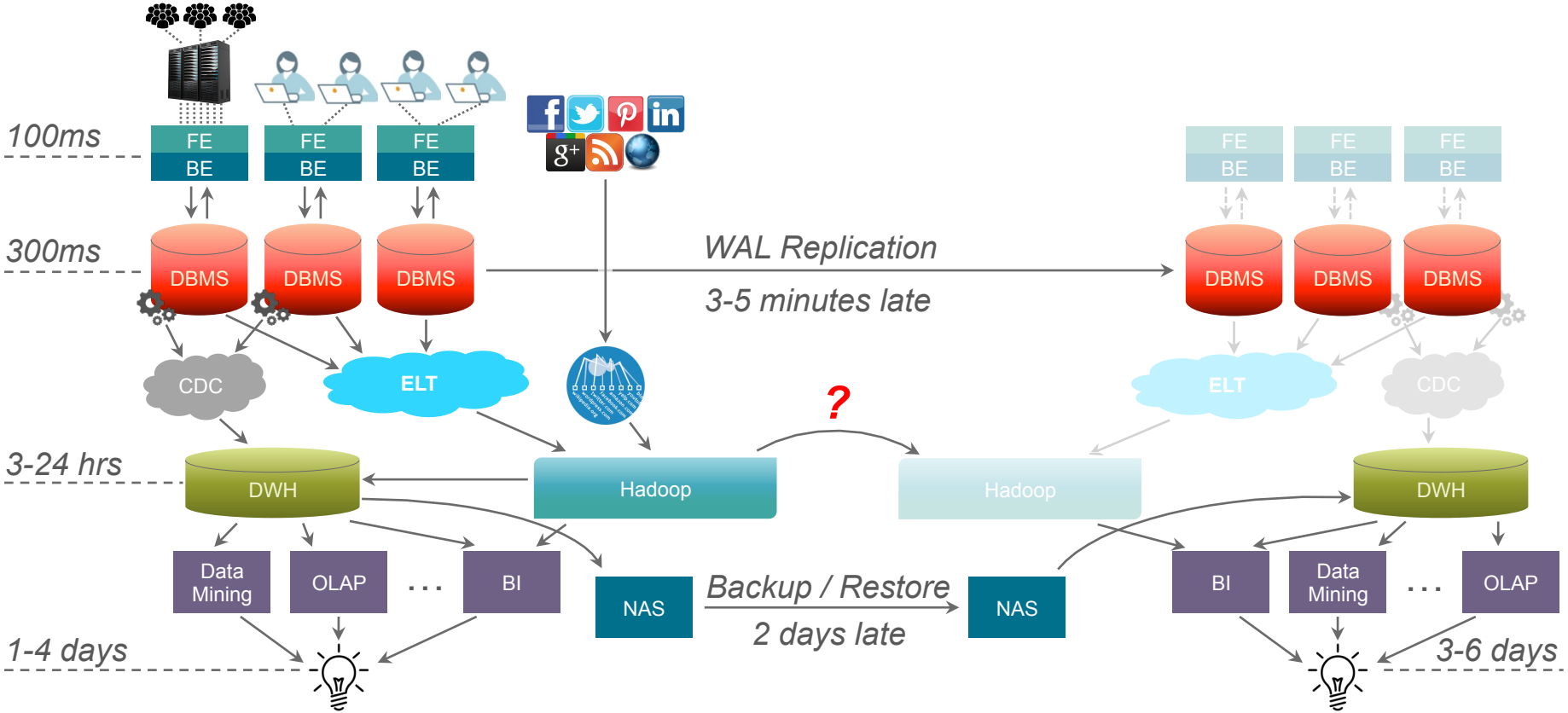
Modern Architectures – Data Lake



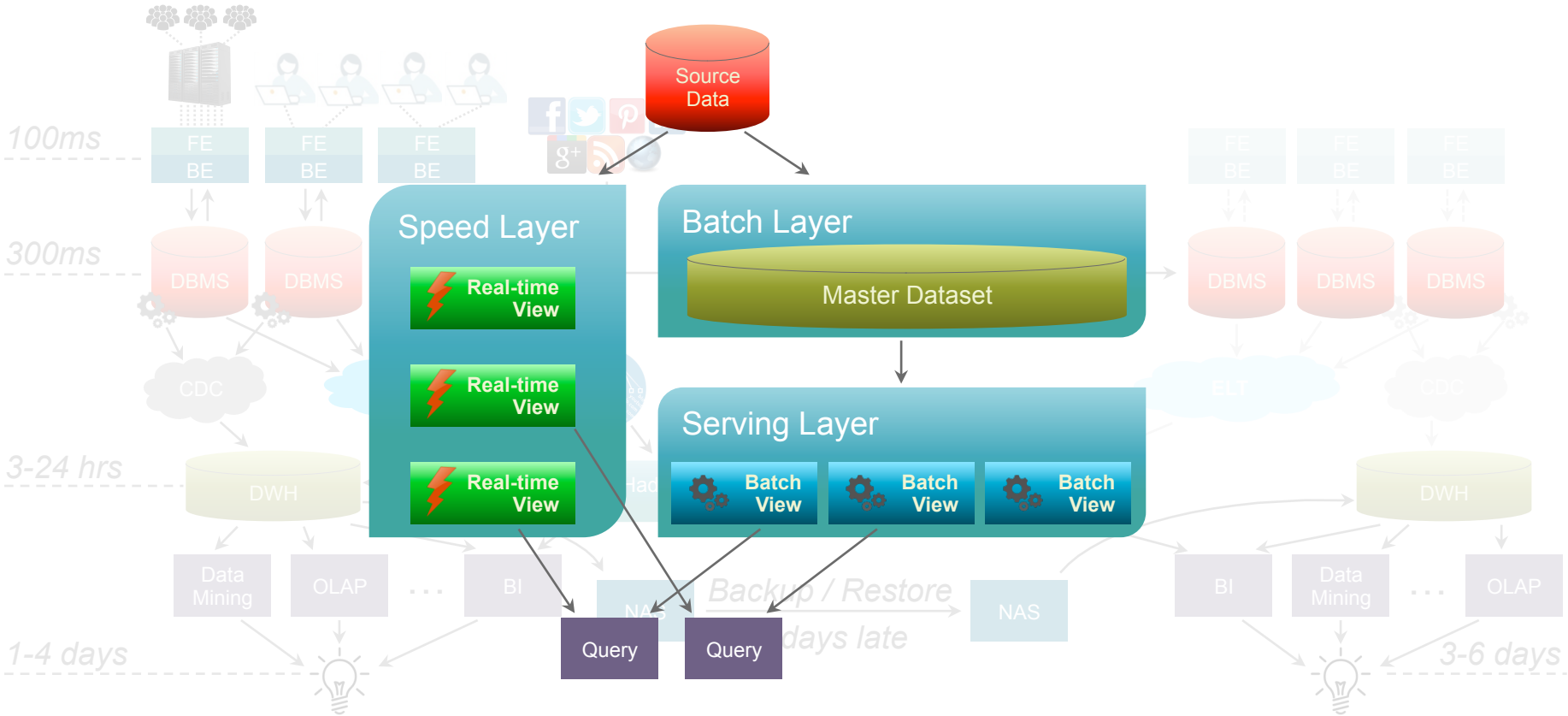
Modern Architectures – Data Lake



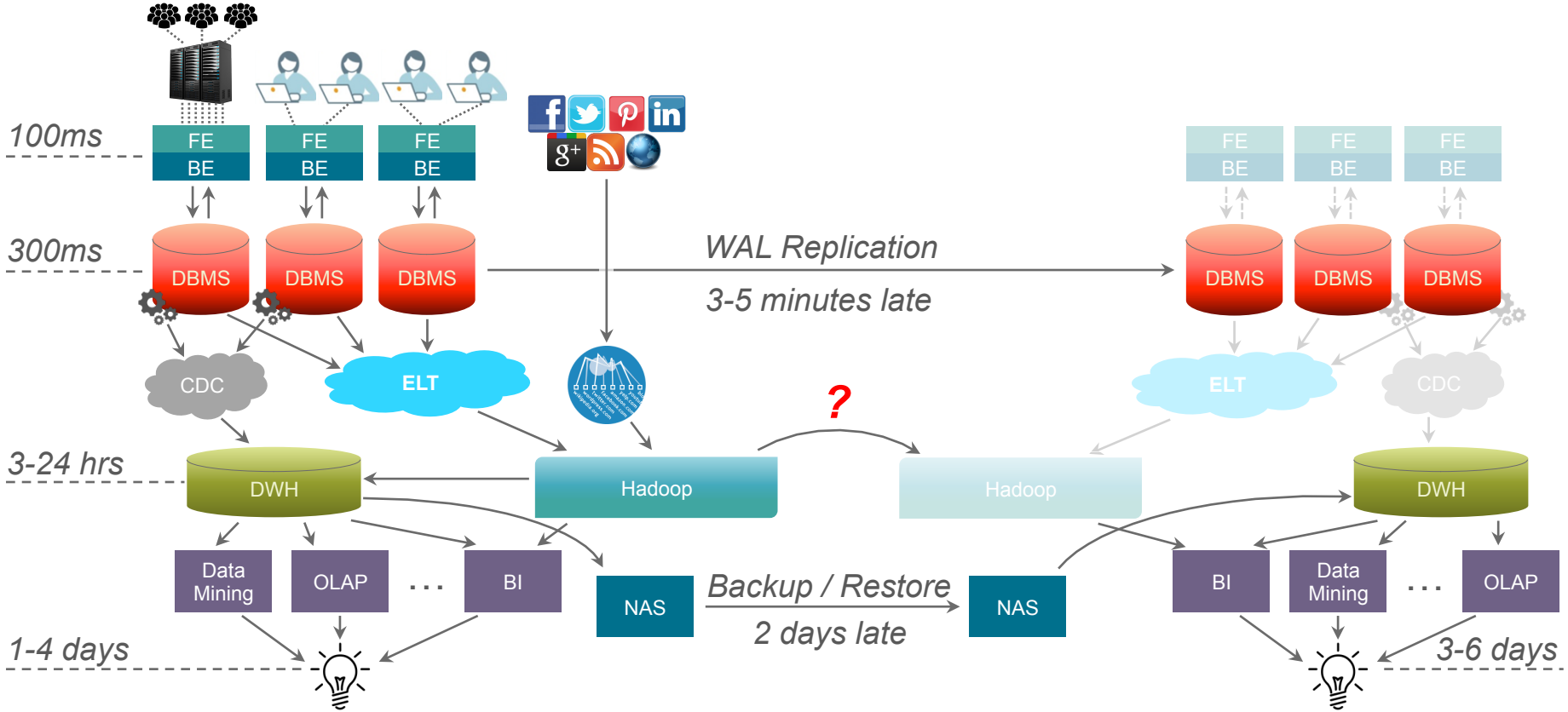
Modern Architectures – Data Lake



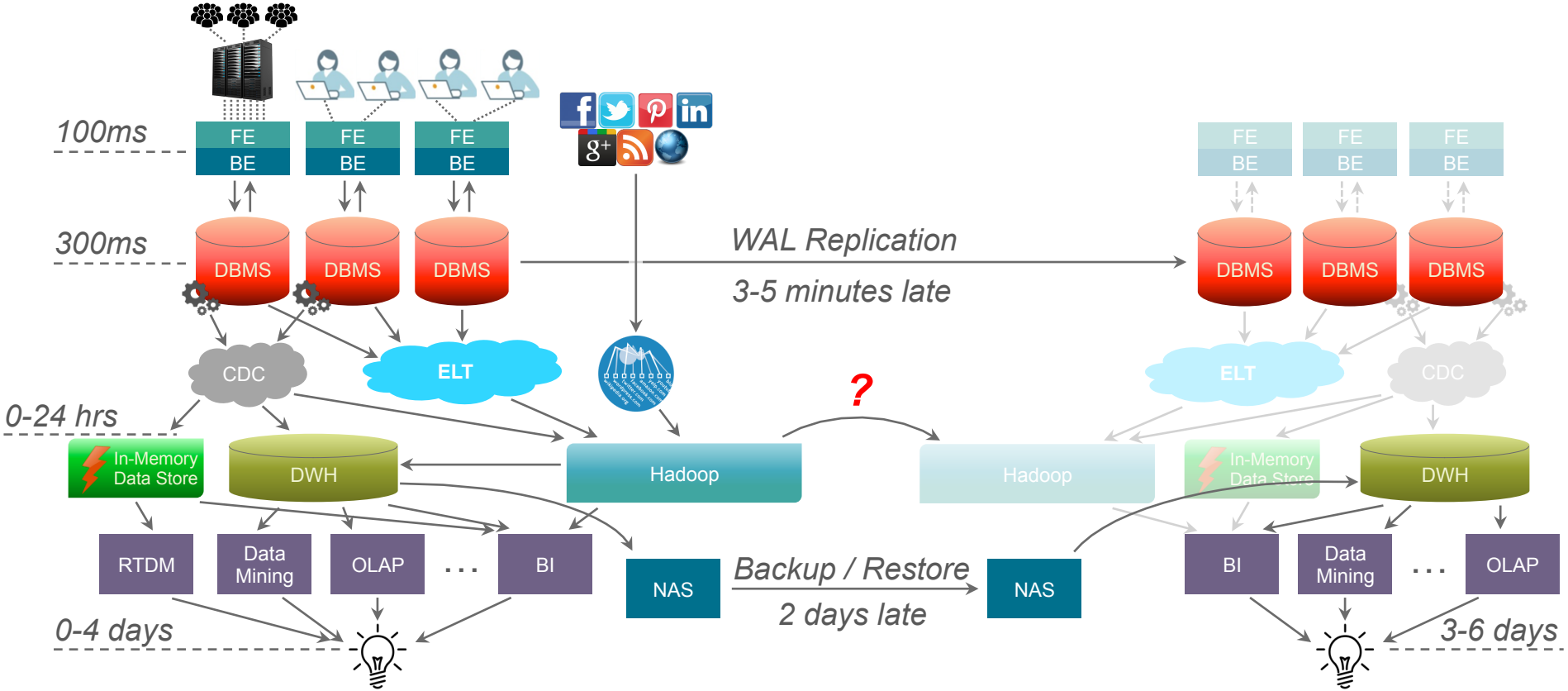
Modern Architectures – Lambda



Modern Architectures – Lambda

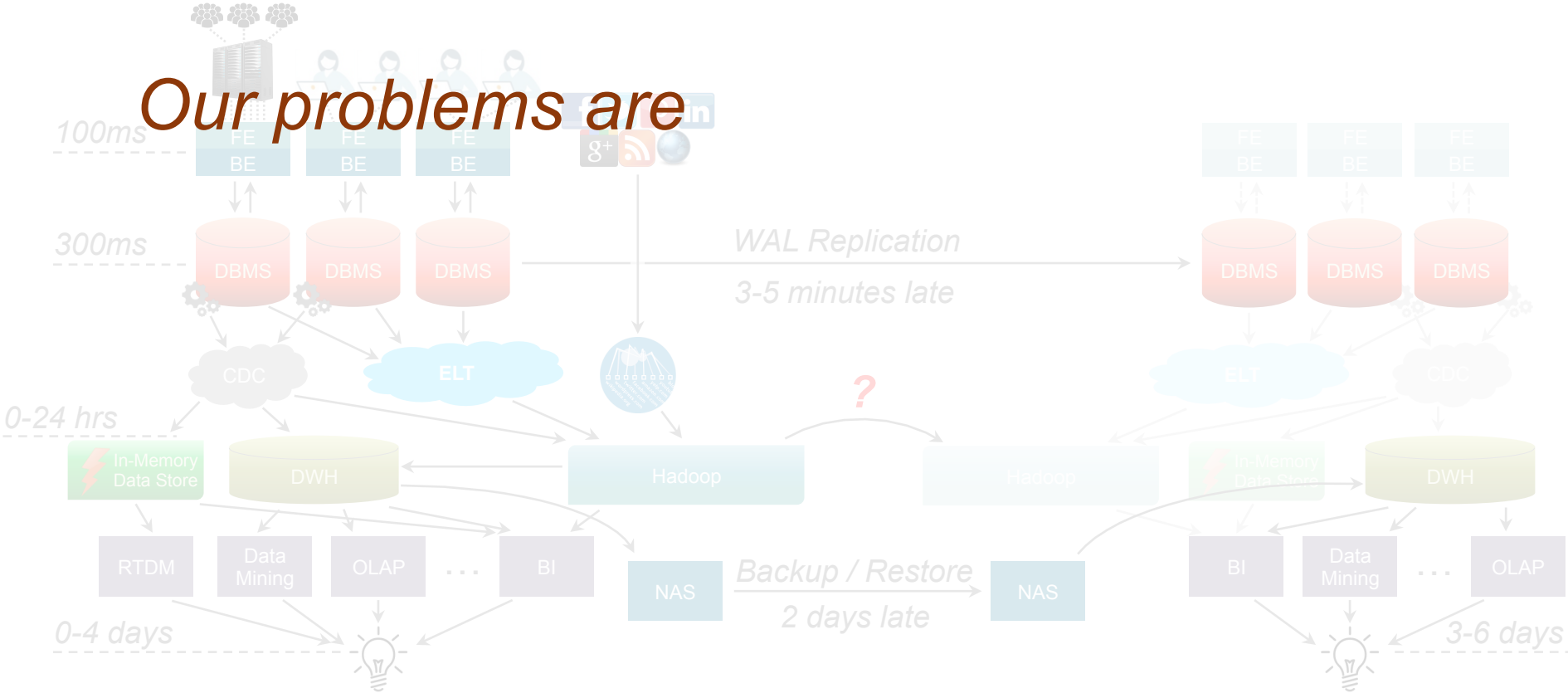


Modern Architectures – Lambda



Modern Architectures

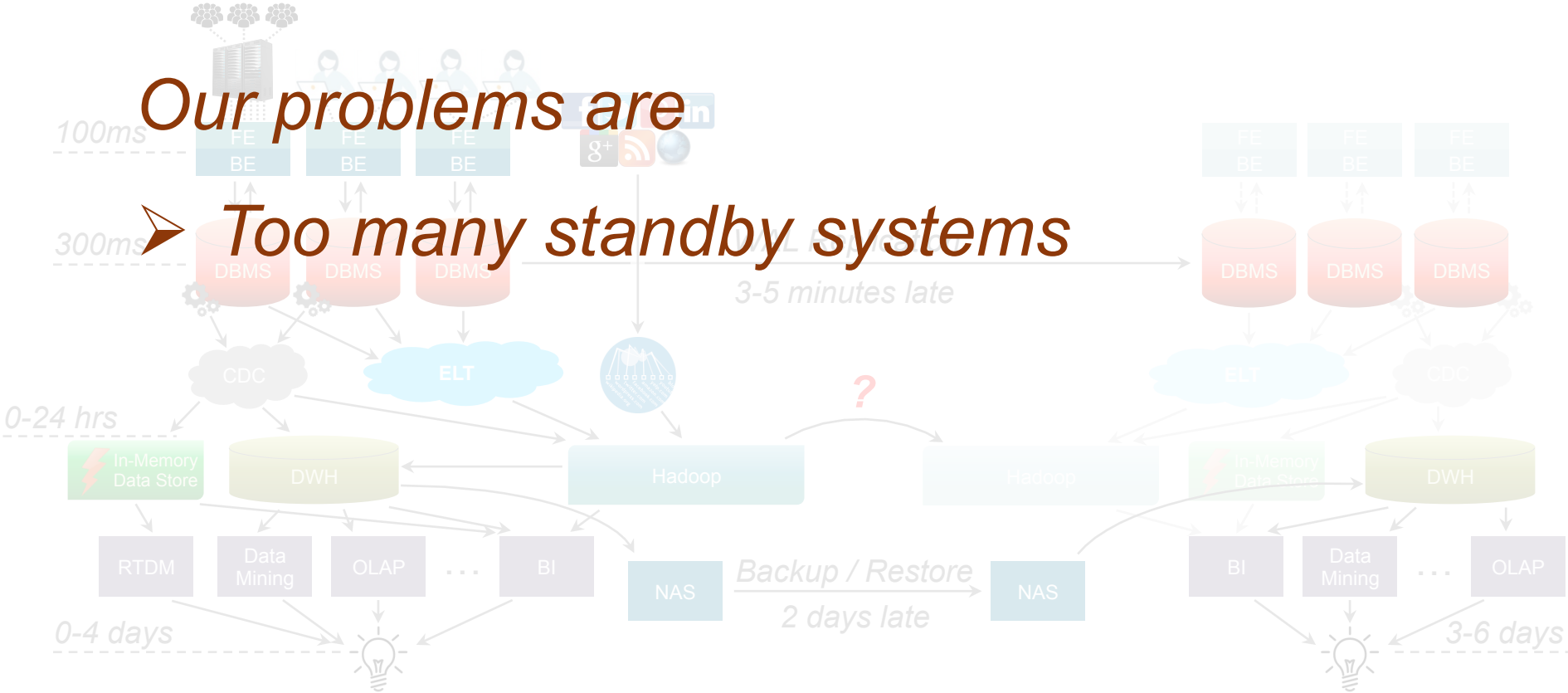
Our problems are



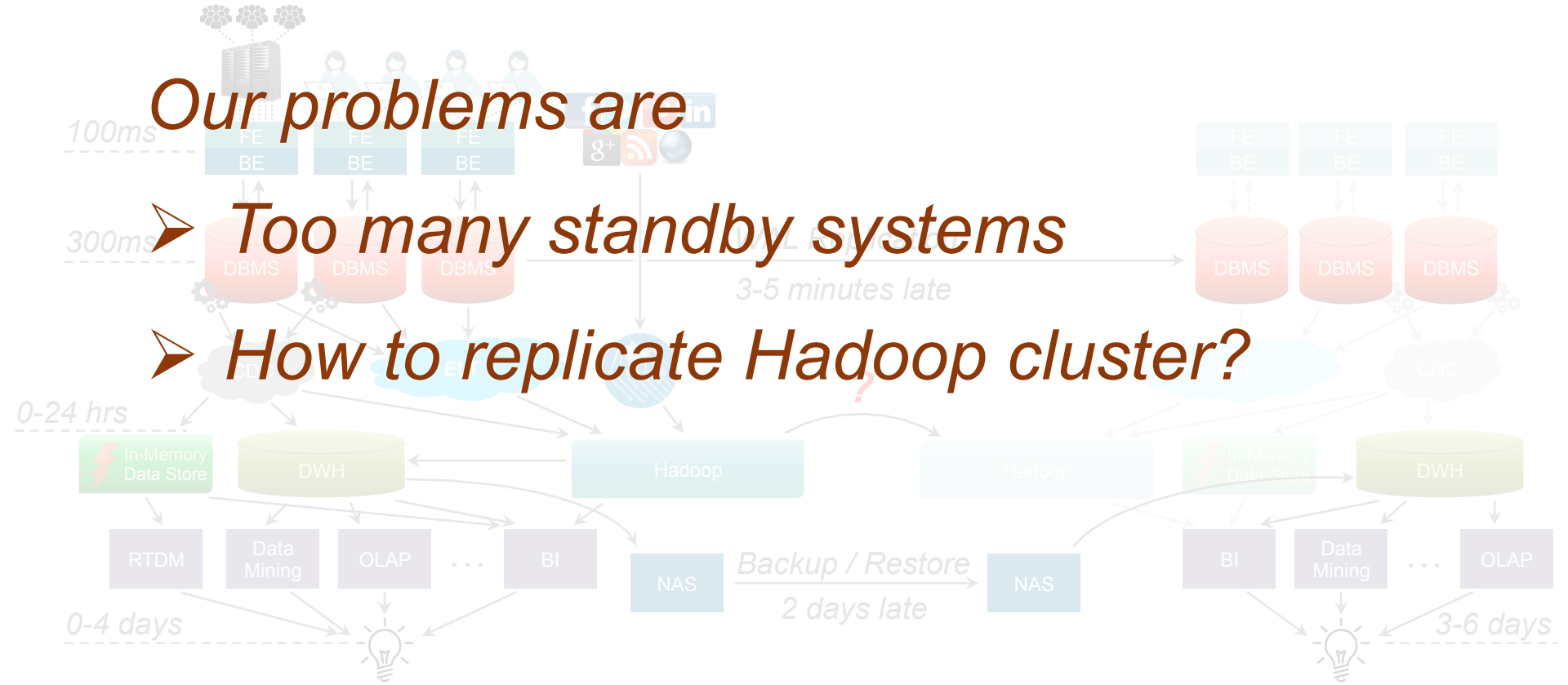
Modern Architectures

Our problems are

- *Too many standby systems*



Modern Architectures



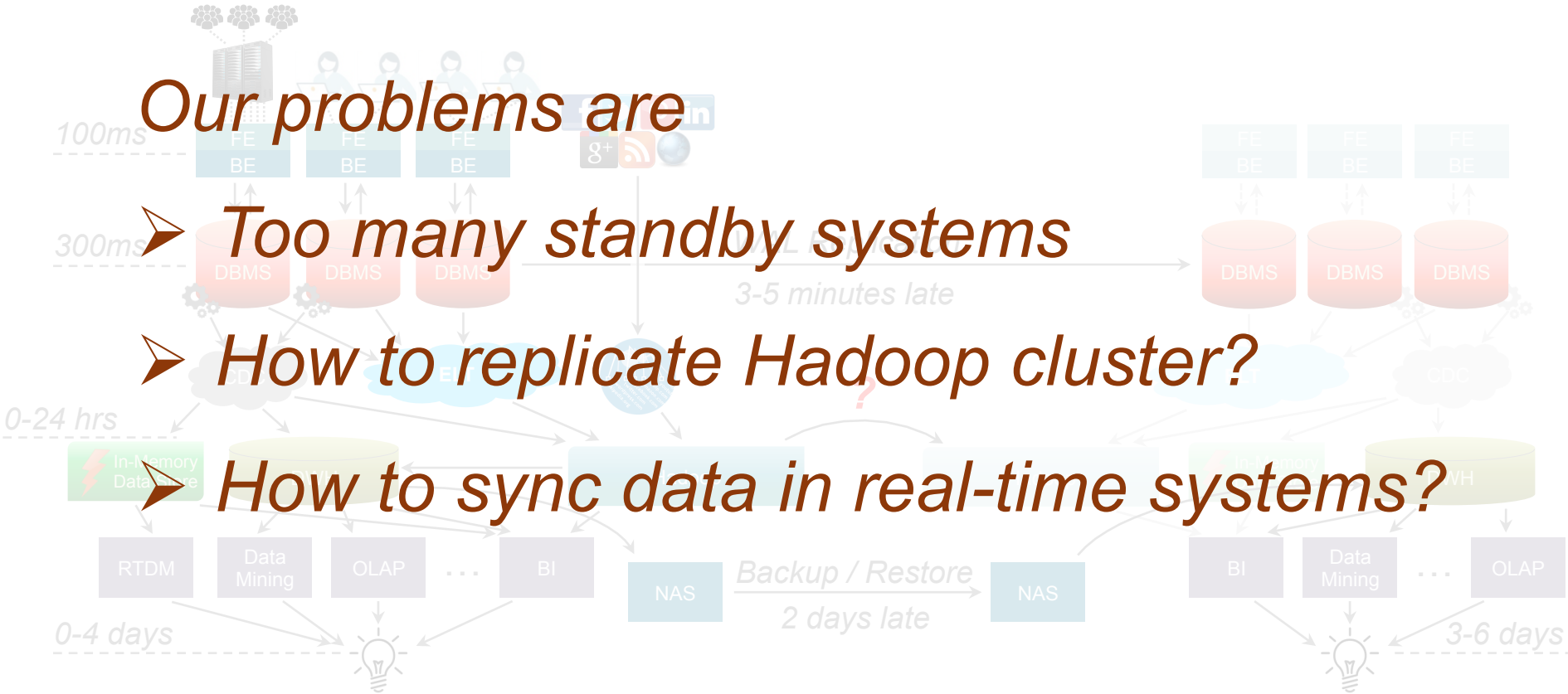
Our problems are

- *Too many standby systems*
- *How to replicate Hadoop cluster?*

Modern Architectures

Our problems are

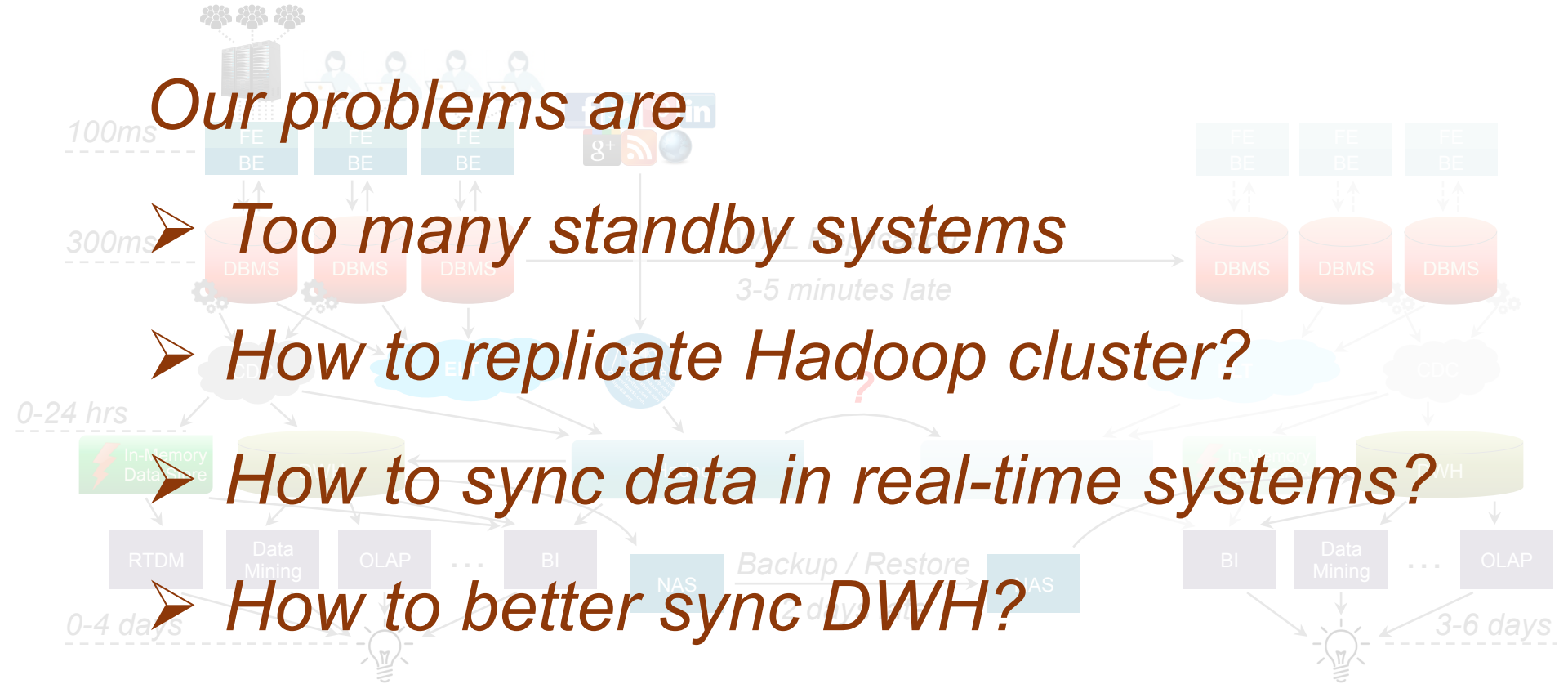
- *Too many standby systems*
- *How to replicate Hadoop cluster?*
- *How to sync data in real-time systems?*



Modern Architectures

Our problems are

- *Too many standby systems*
- *How to replicate Hadoop cluster?*
- *How to sync data in real-time systems?*
- *How to better sync DWH?*

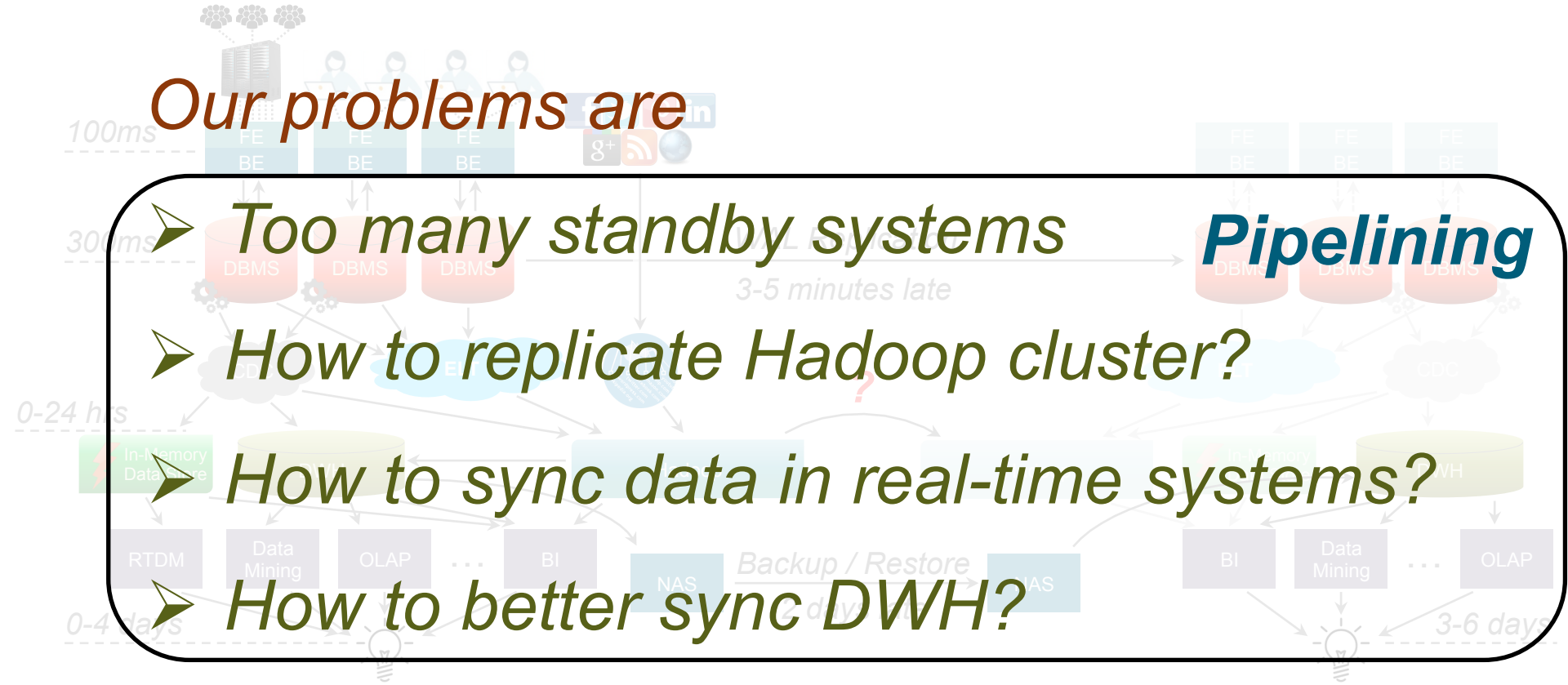


Modern Architectures

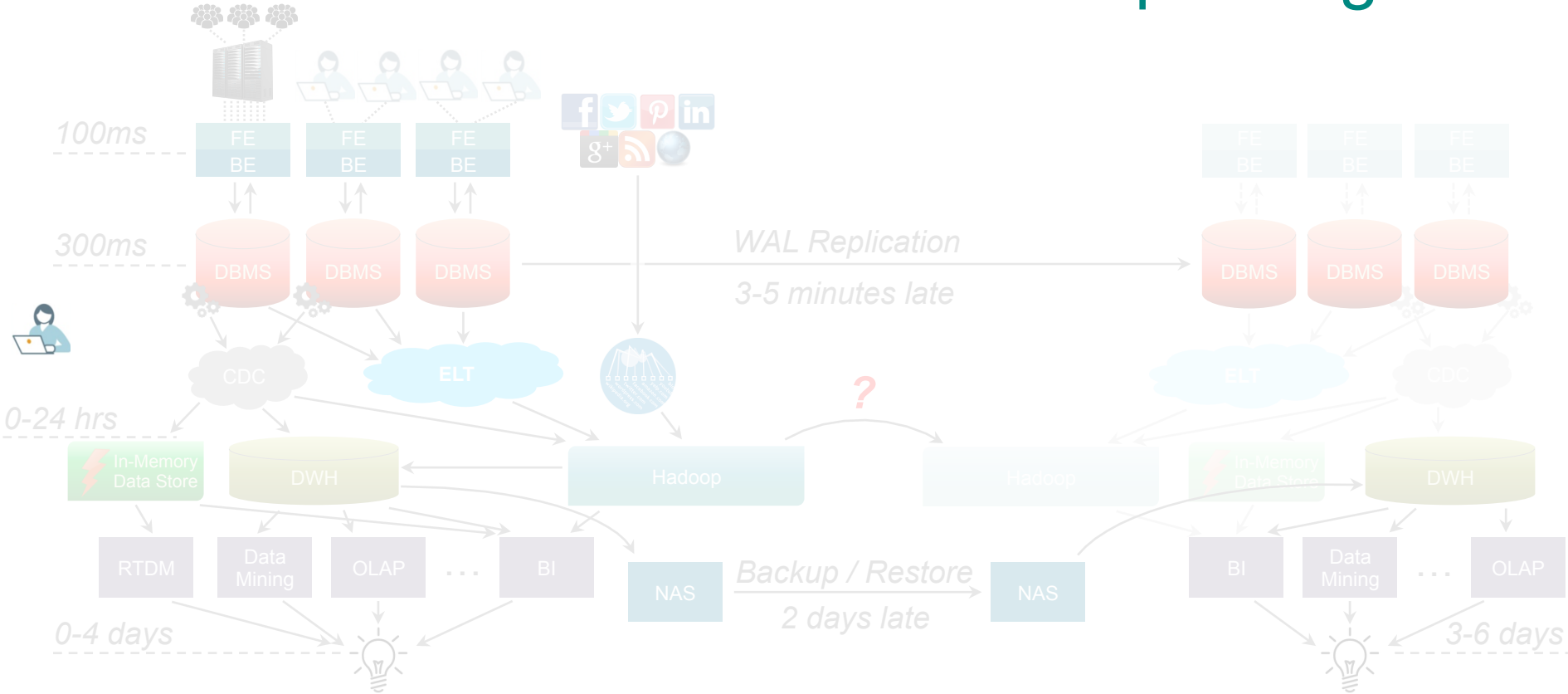
Our problems are

- *Too many standby systems*
- *How to replicate Hadoop cluster?*
- *How to sync data in real-time systems?*
- *How to better sync DWH?*

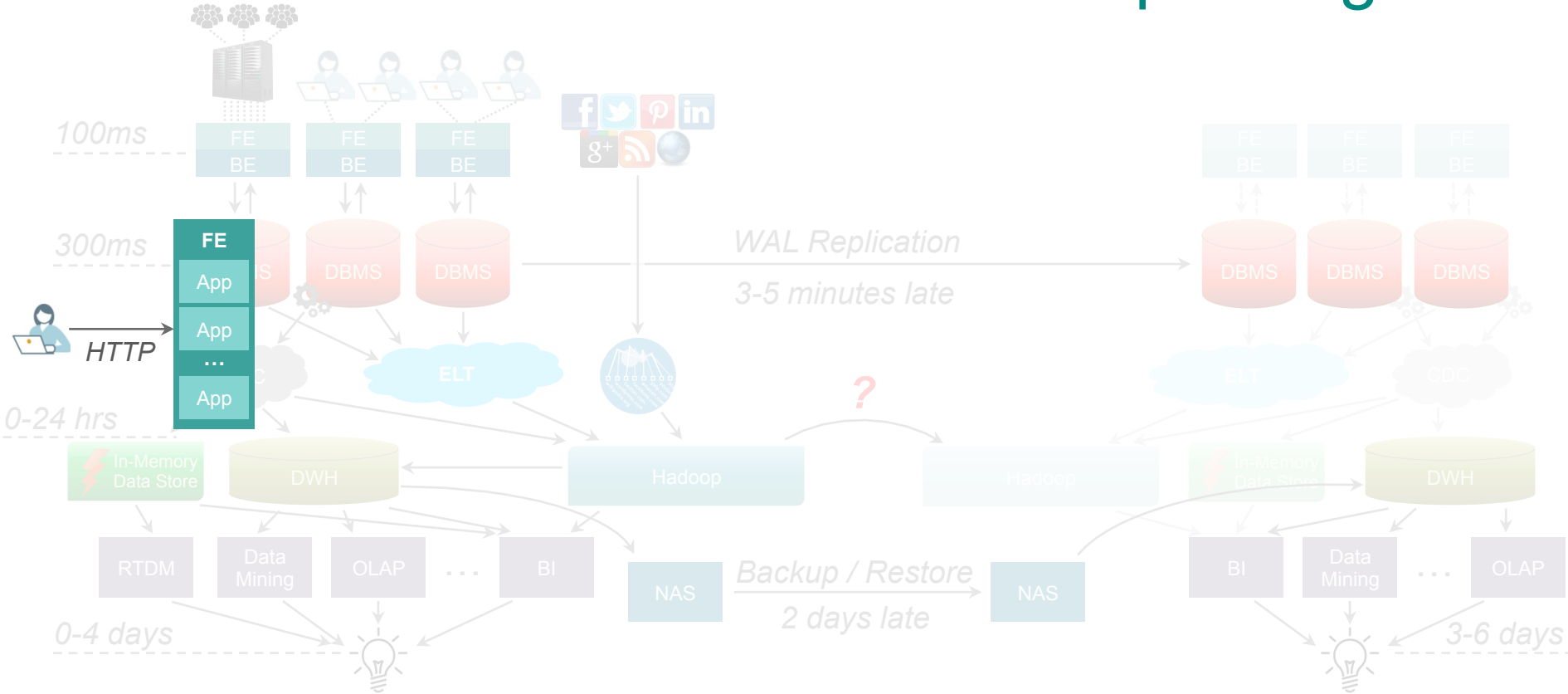
Pipelining



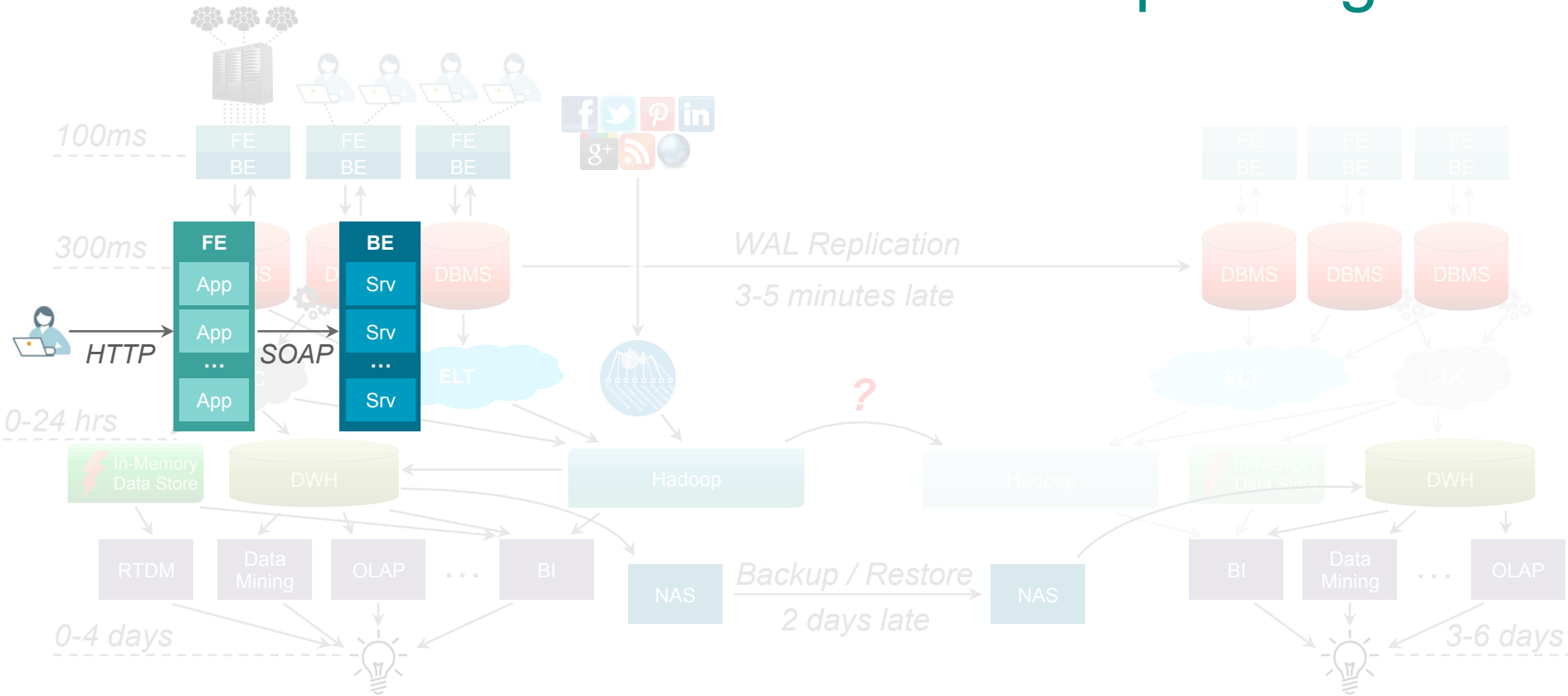
Modern Data Architecture – Pipelining



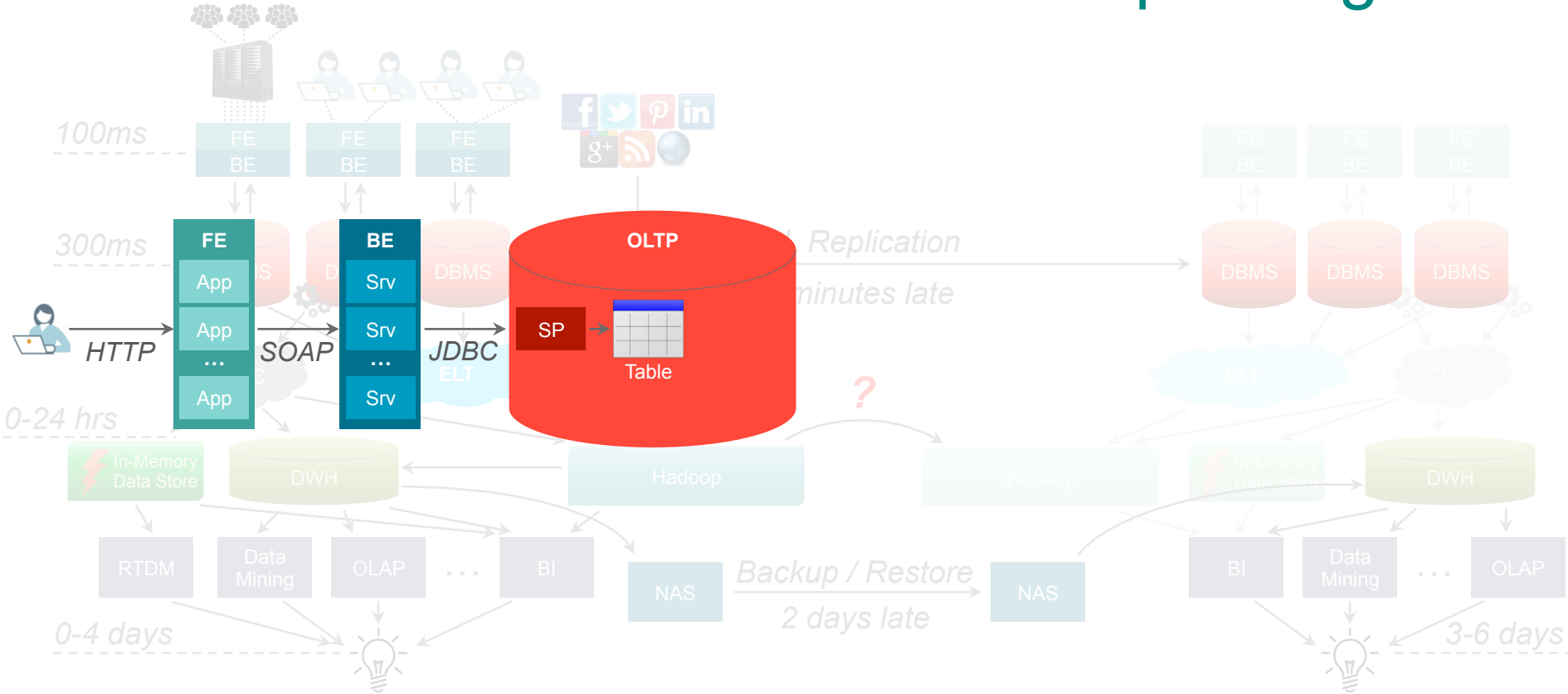
Modern Data Architecture – Pipelining



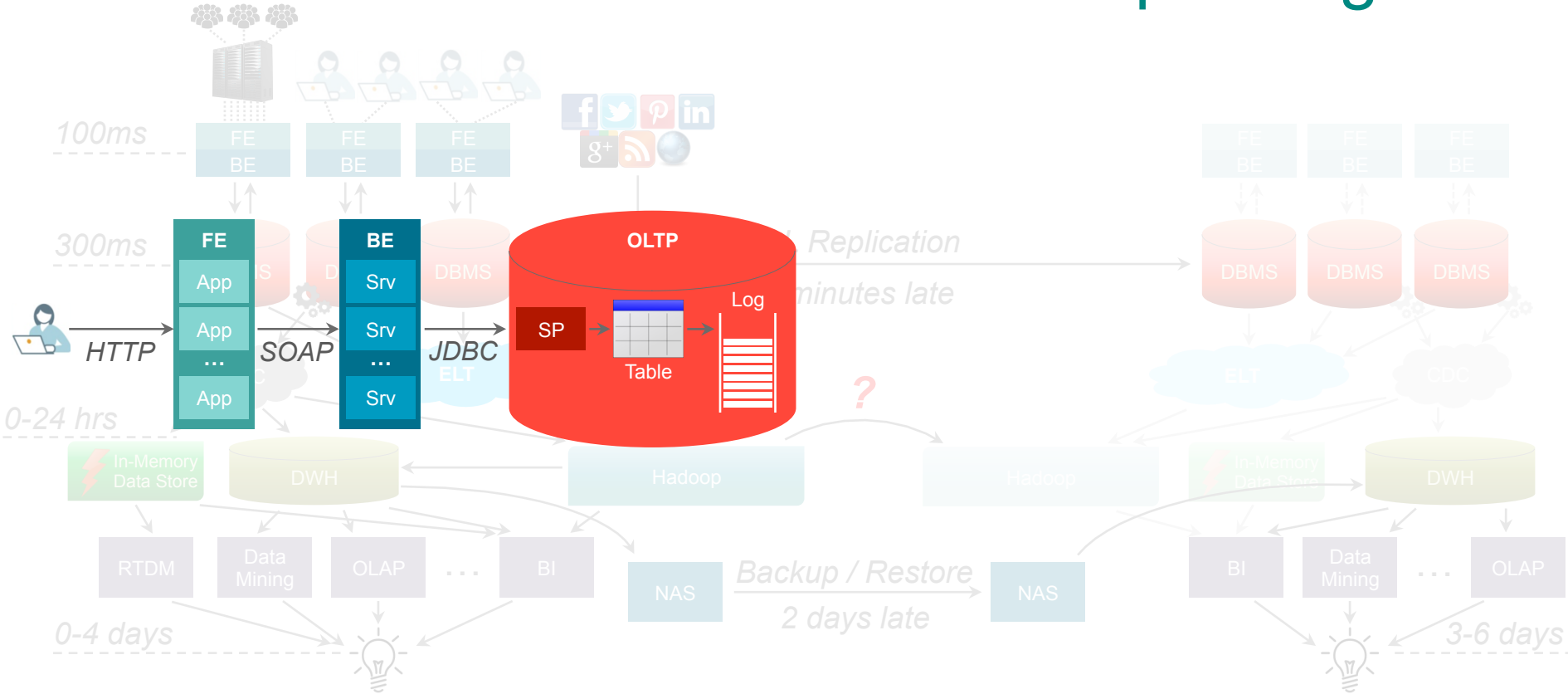
Modern Data Architecture – Pipelining



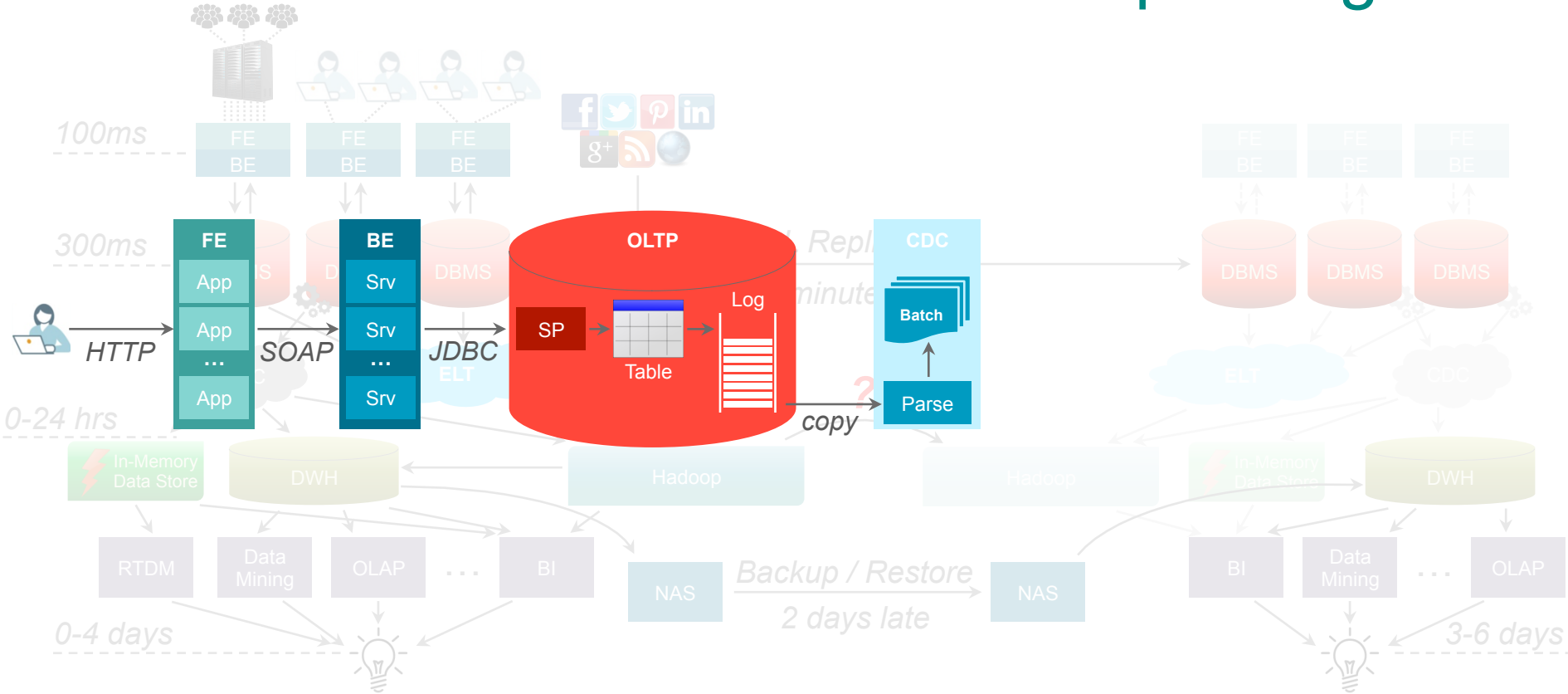
Modern Data Architecture – Pipelining



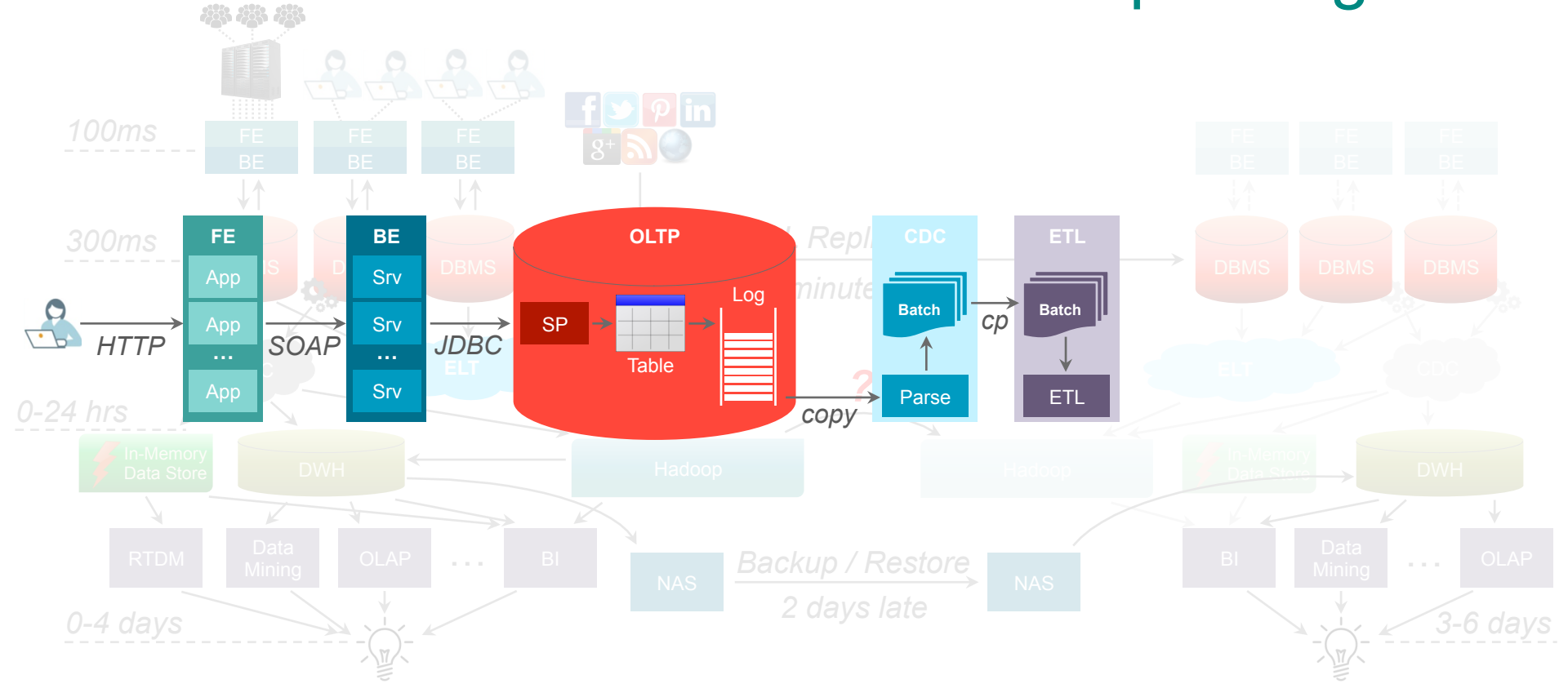
Modern Data Architecture – Pipelining



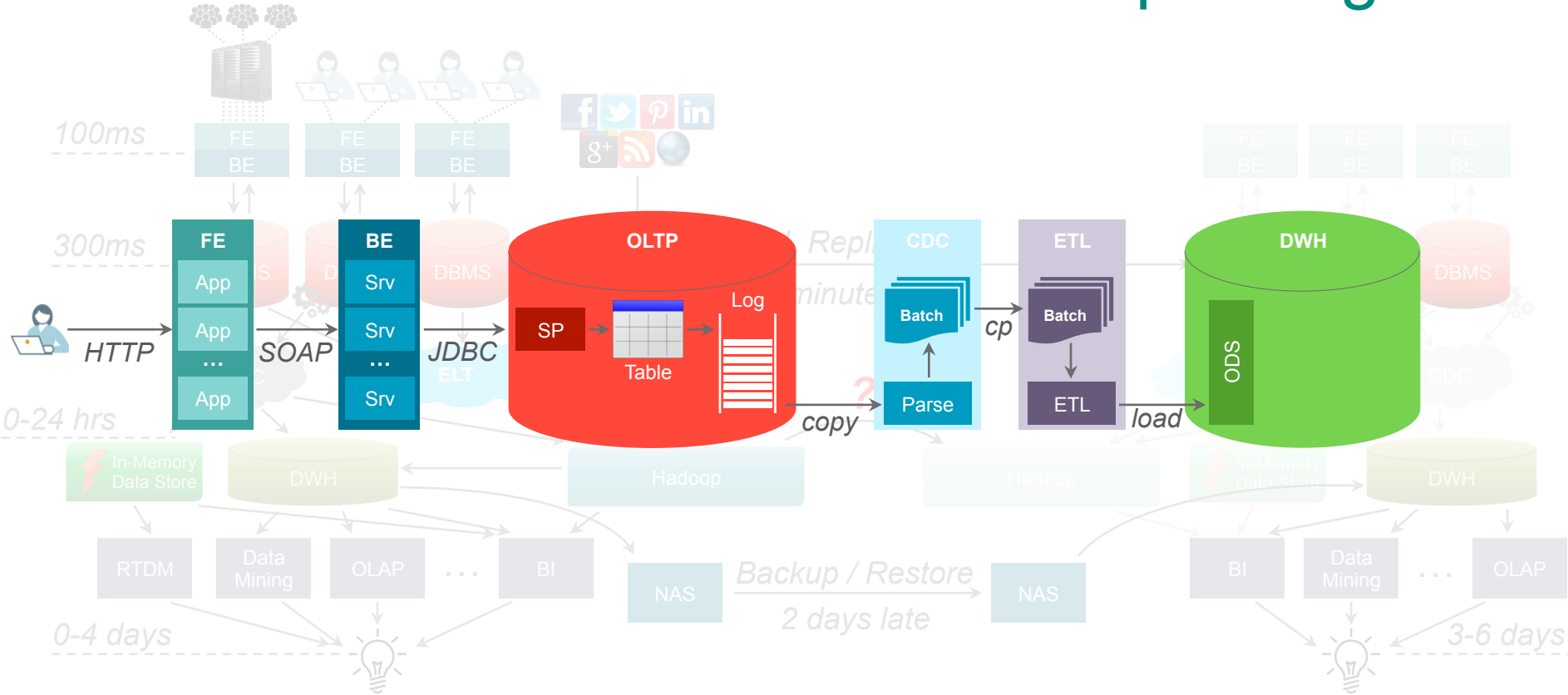
Modern Data Architecture – Pipelining



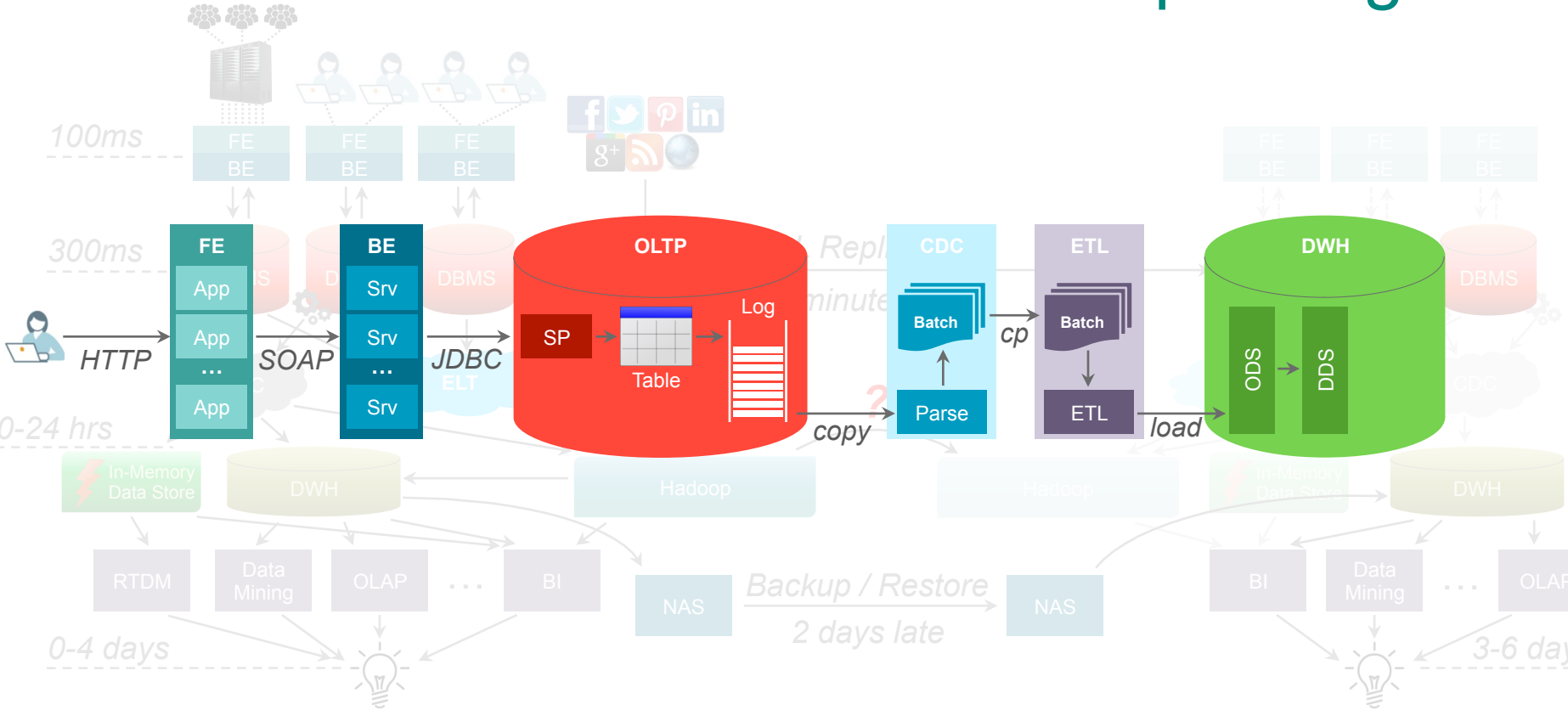
Modern Data Architecture – Pipelining



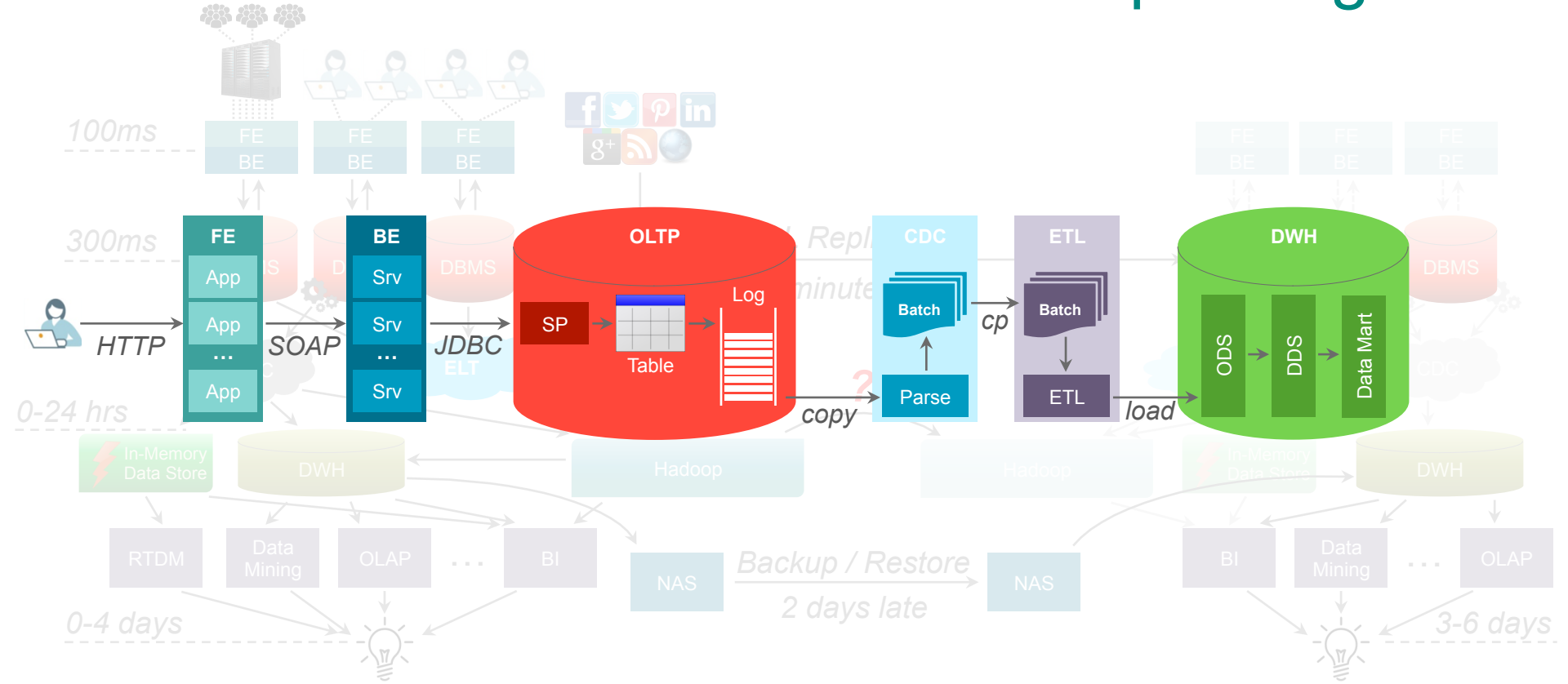
Modern Data Architecture – Pipelining



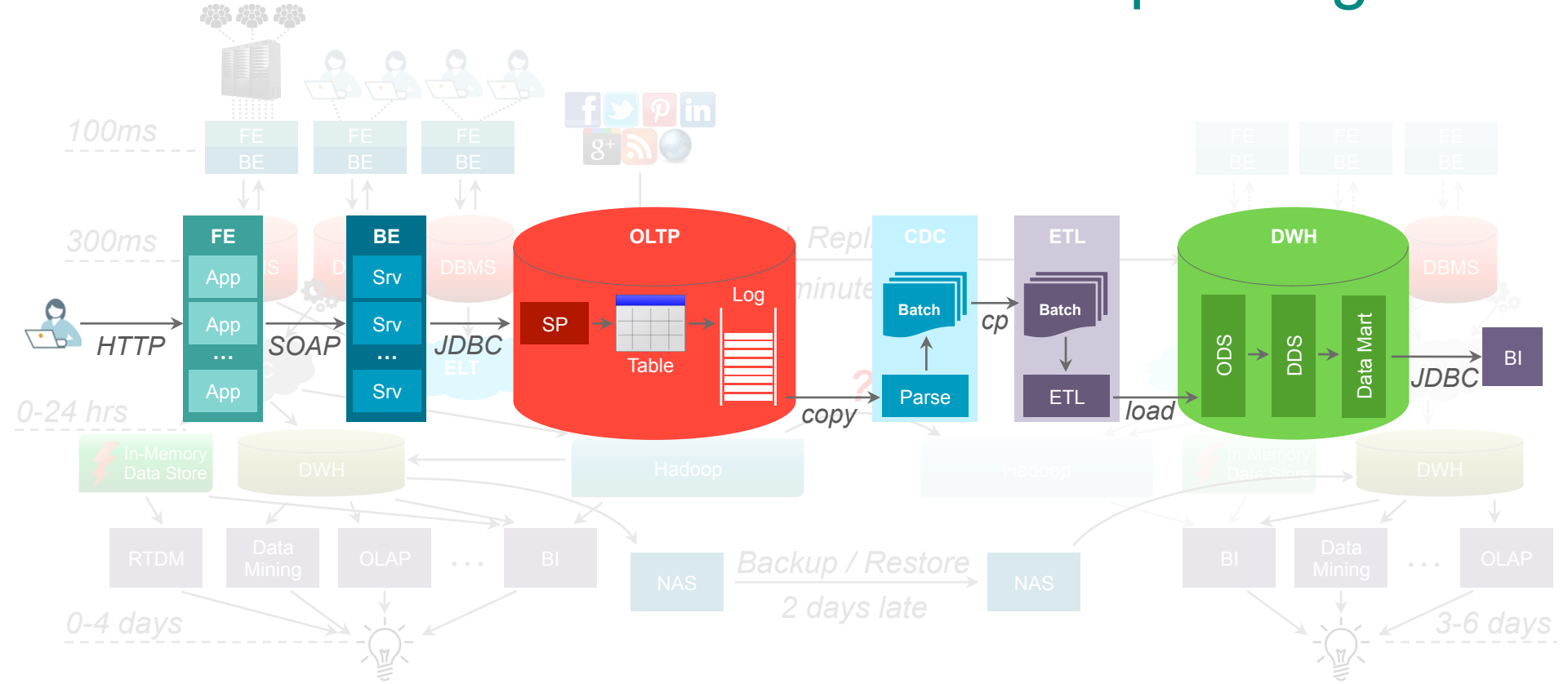
Modern Data Architecture – Pipelining



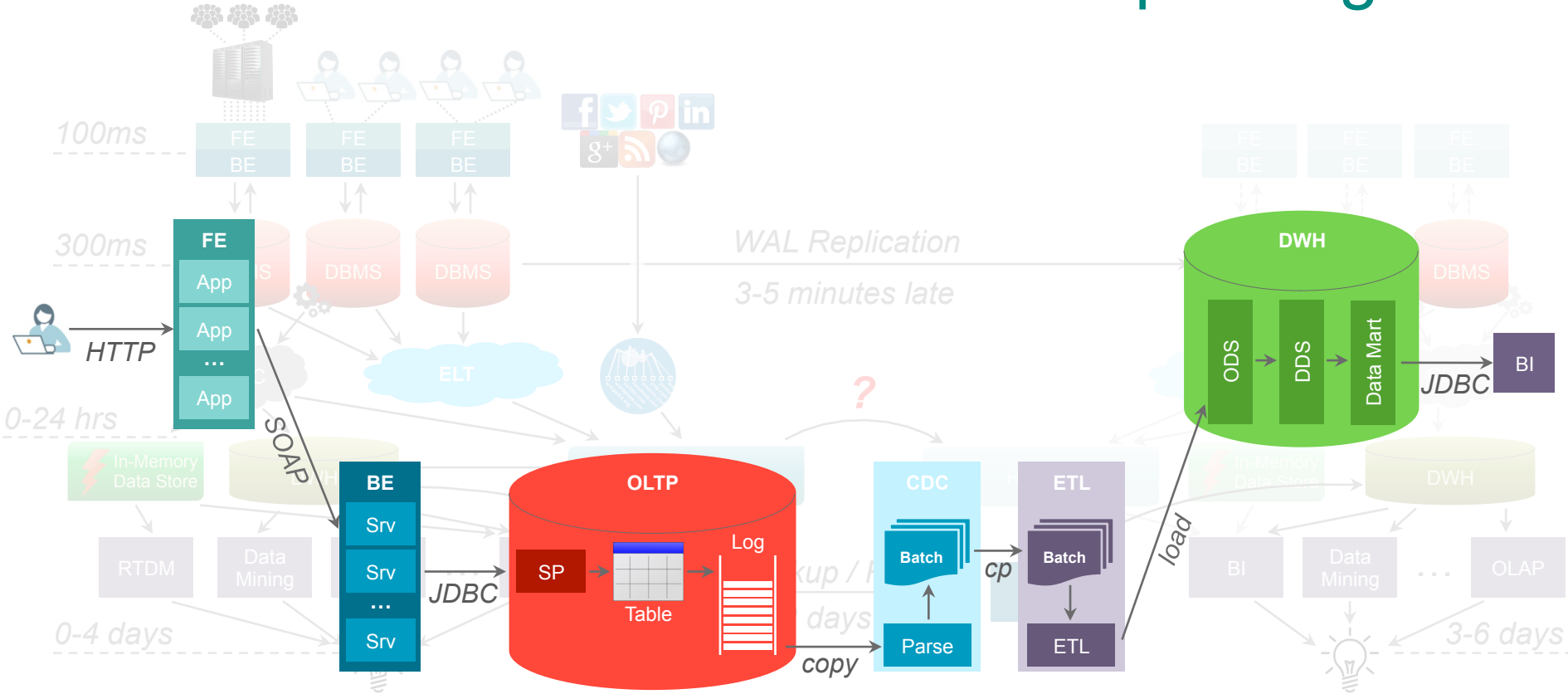
Modern Data Architecture – Pipelining



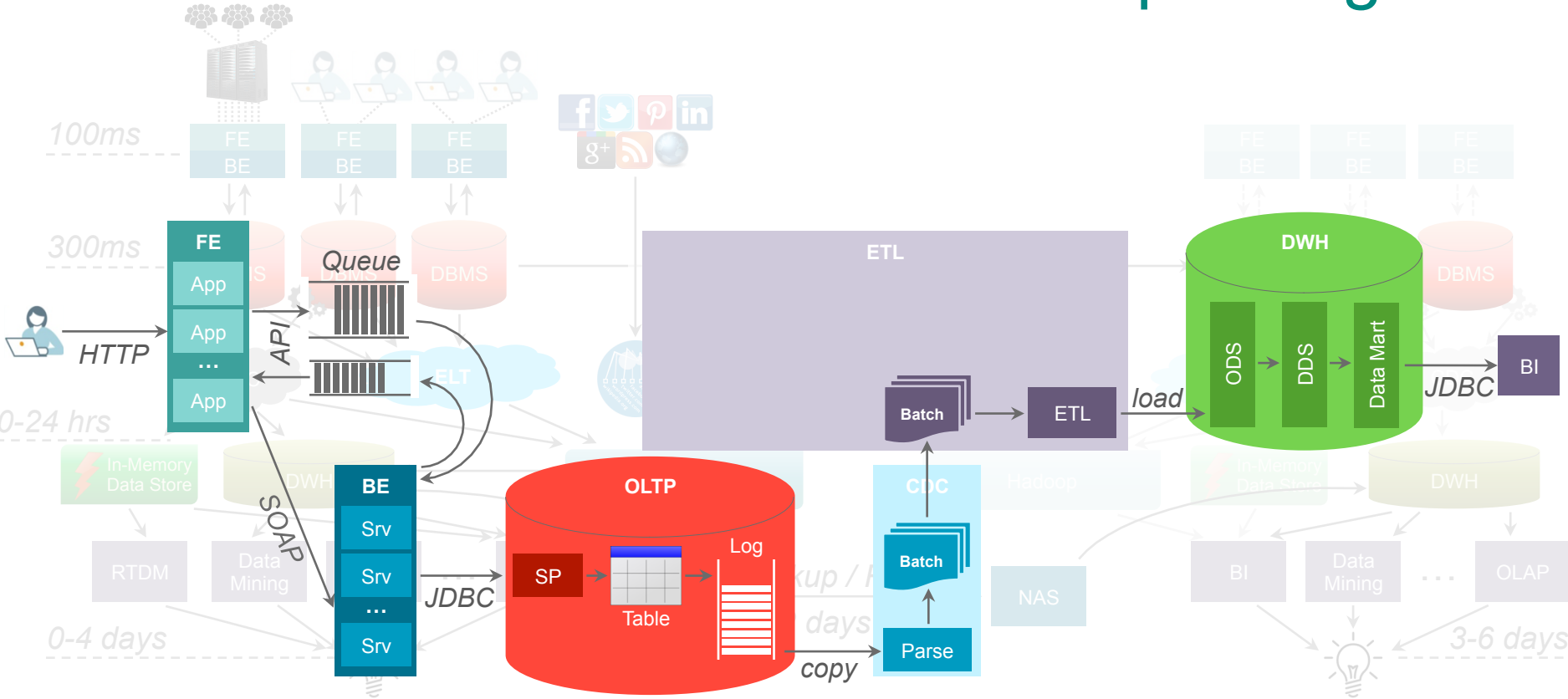
Modern Data Architecture – Pipelining



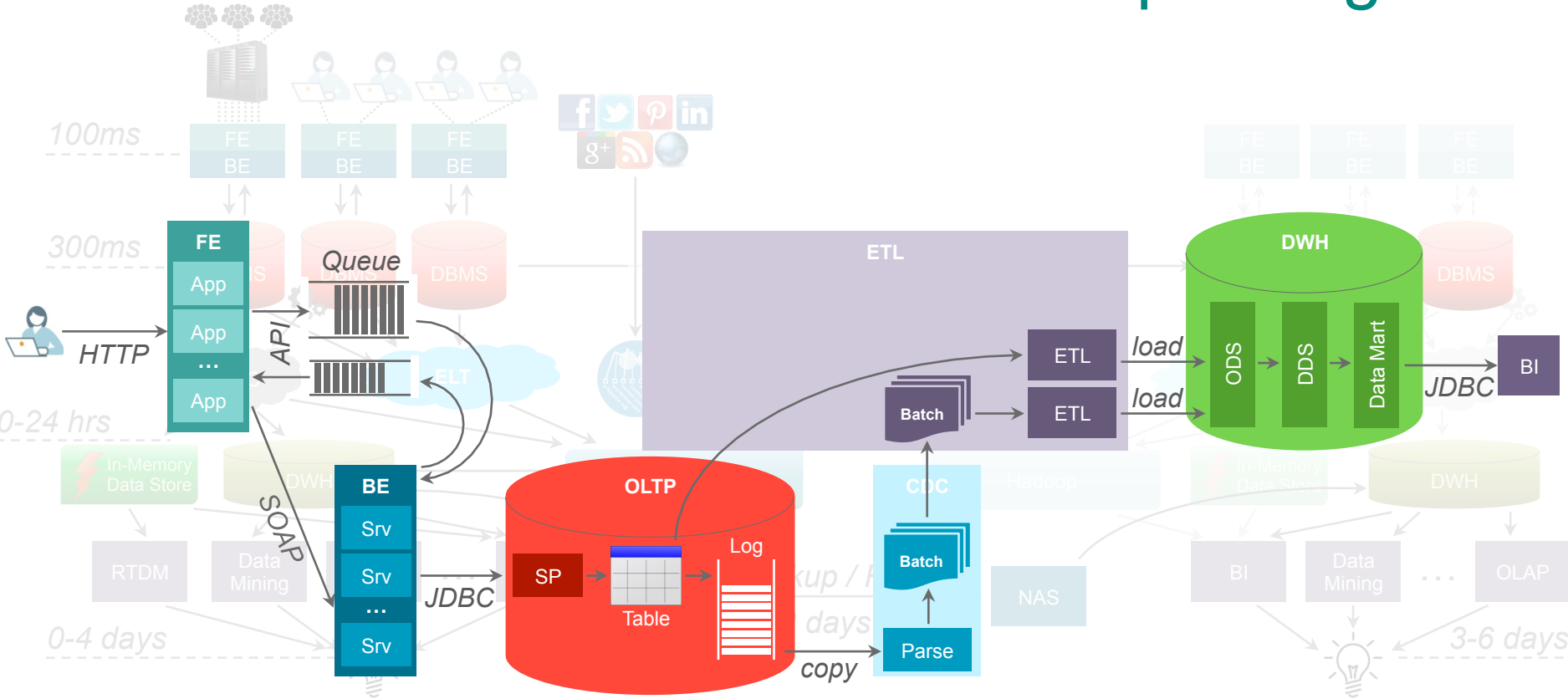
Modern Data Architecture – Pipelining



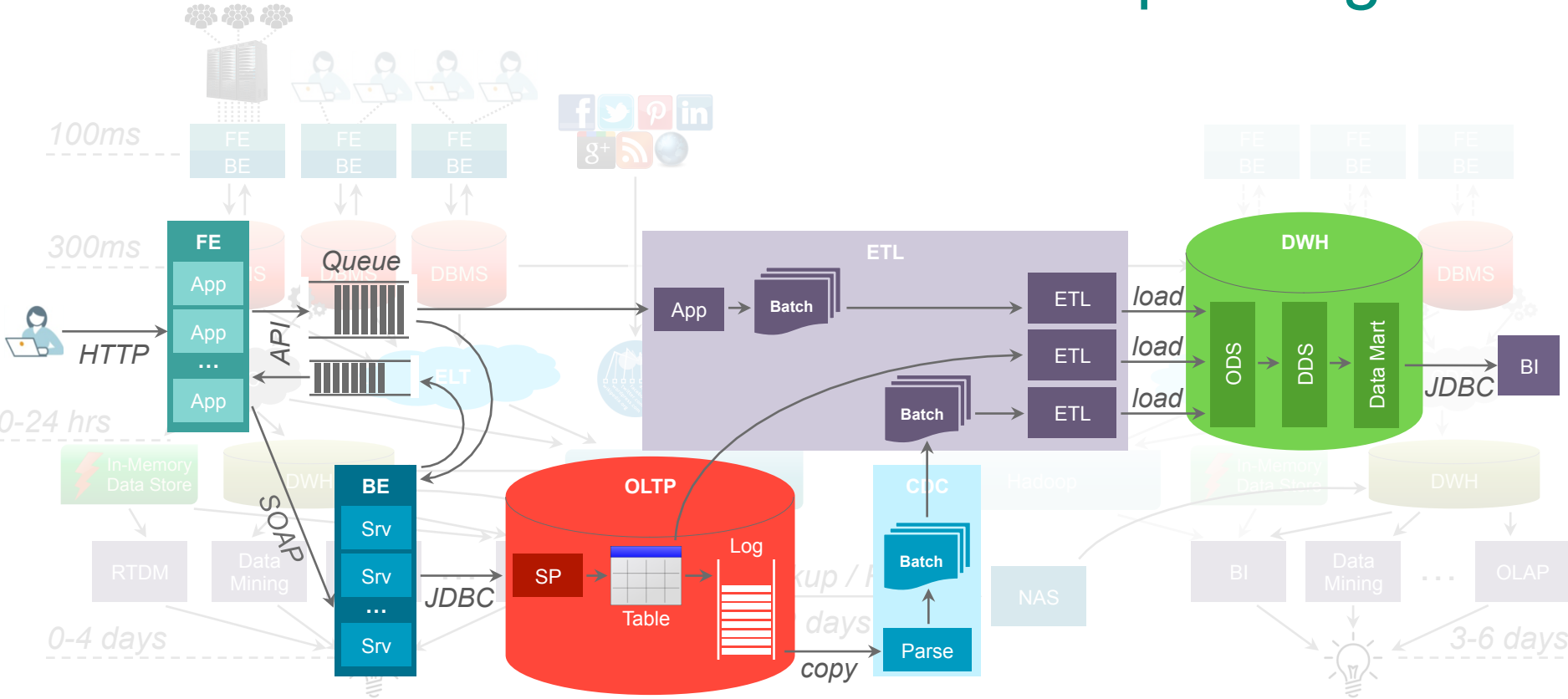
Modern Data Architecture – Pipelining



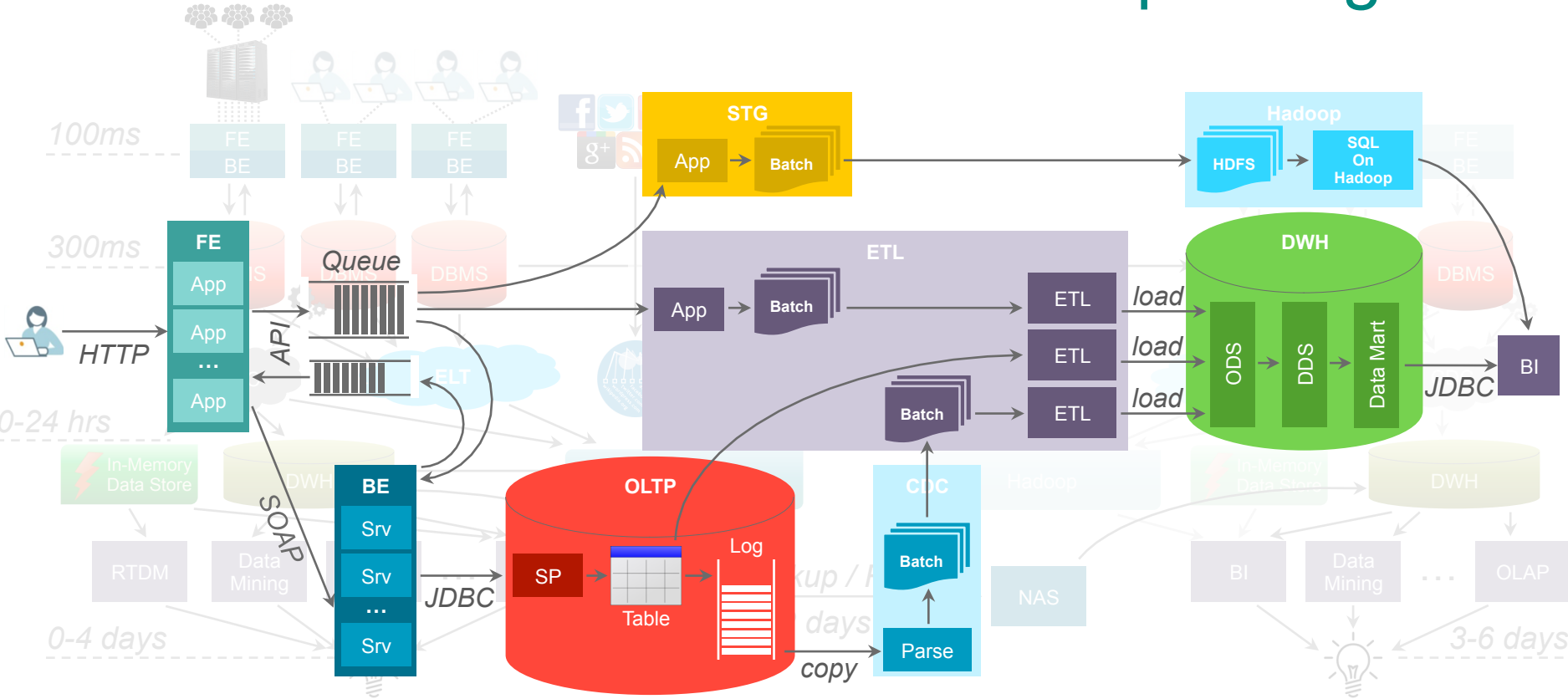
Modern Data Architecture – Pipelining



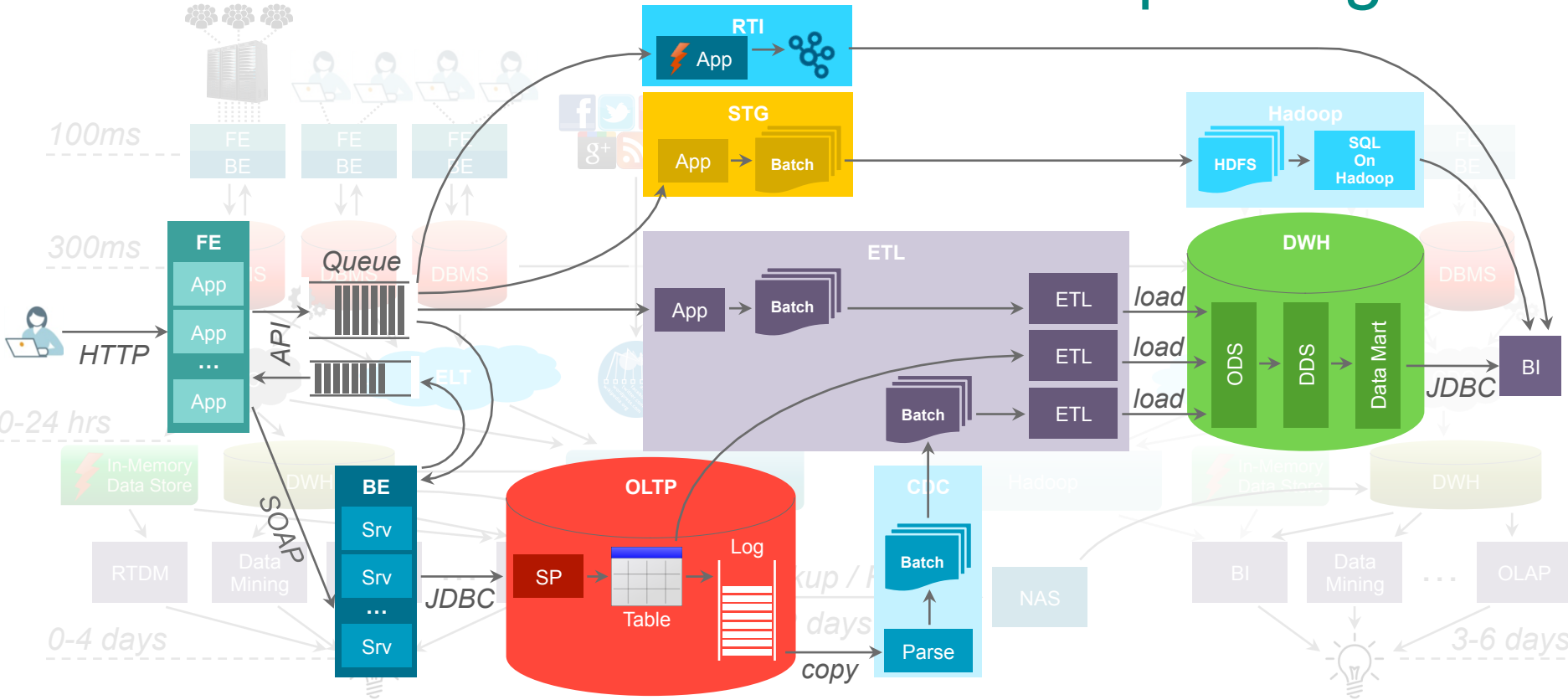
Modern Data Architecture – Pipelining



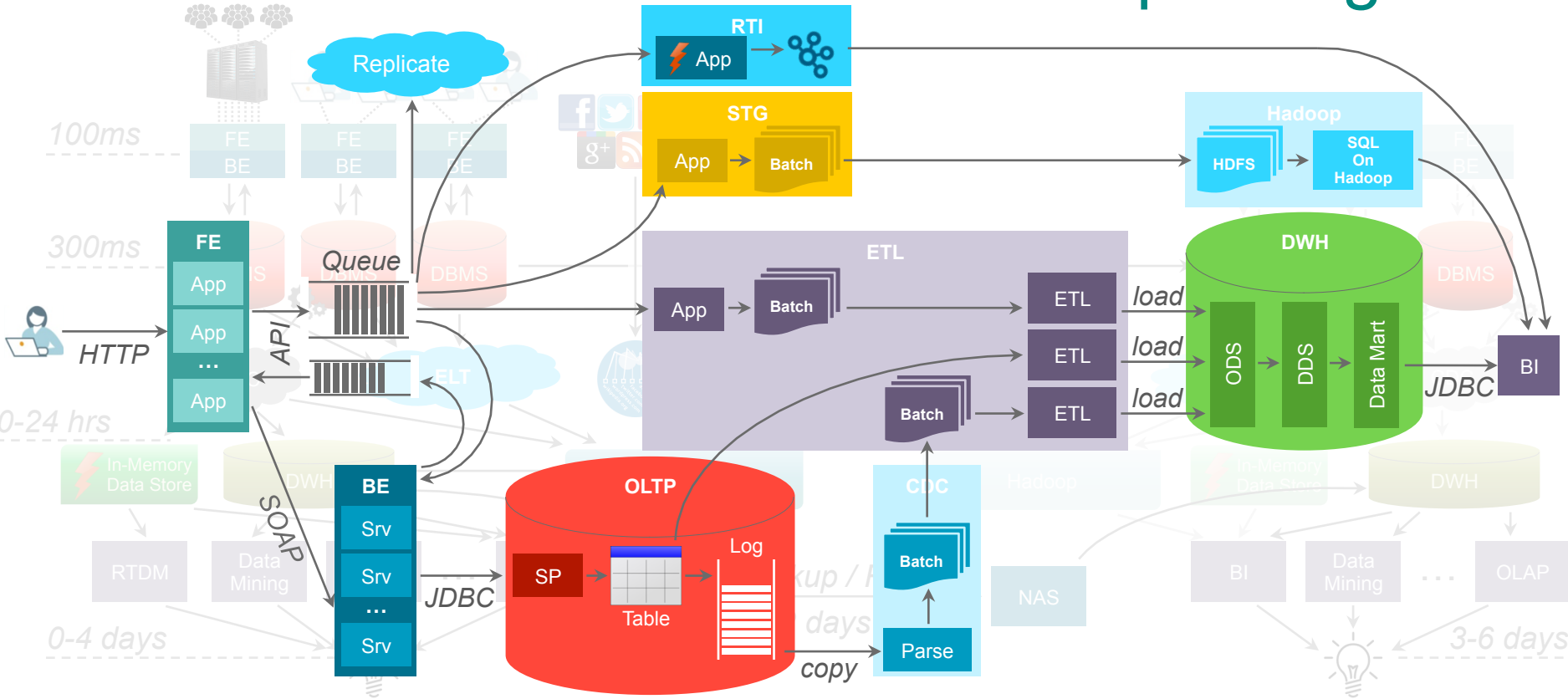
Modern Data Architecture – Pipelining



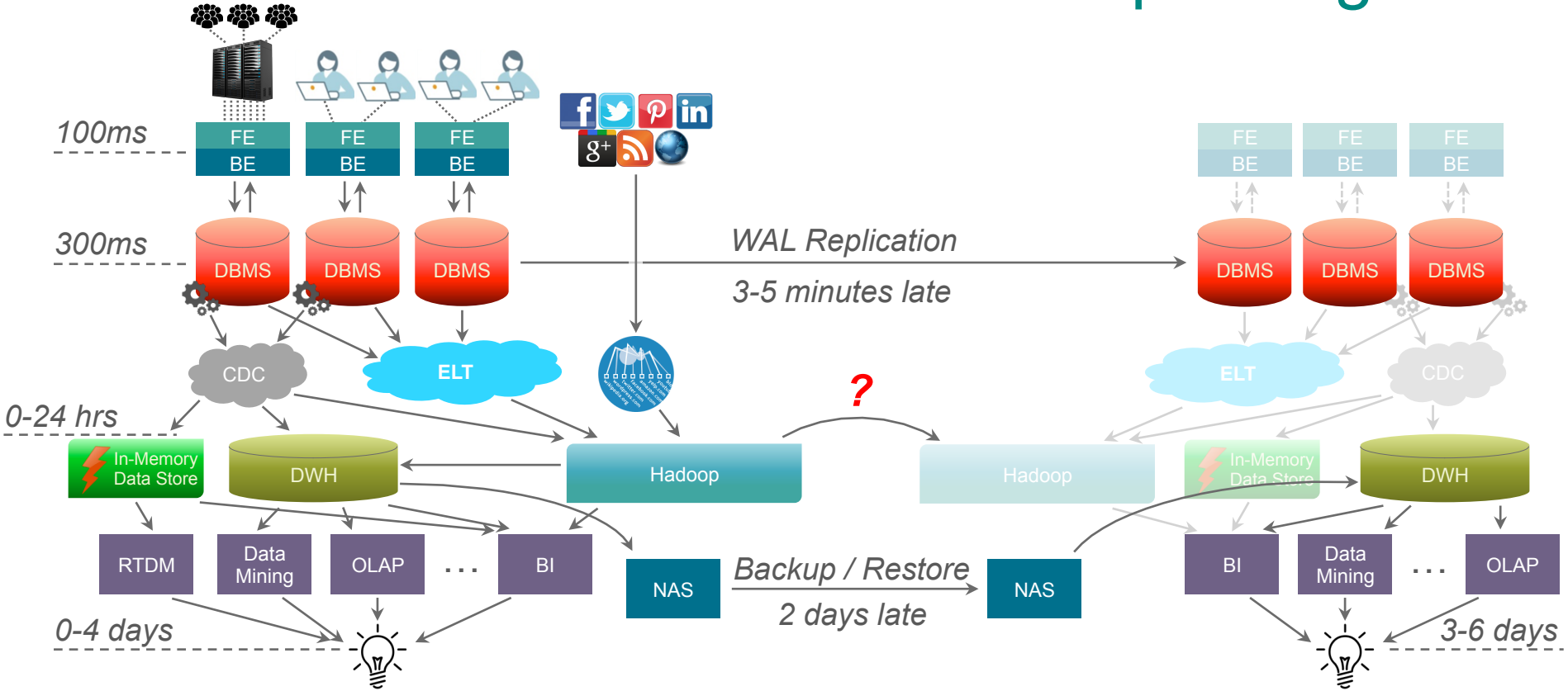
Modern Data Architecture – Pipelining



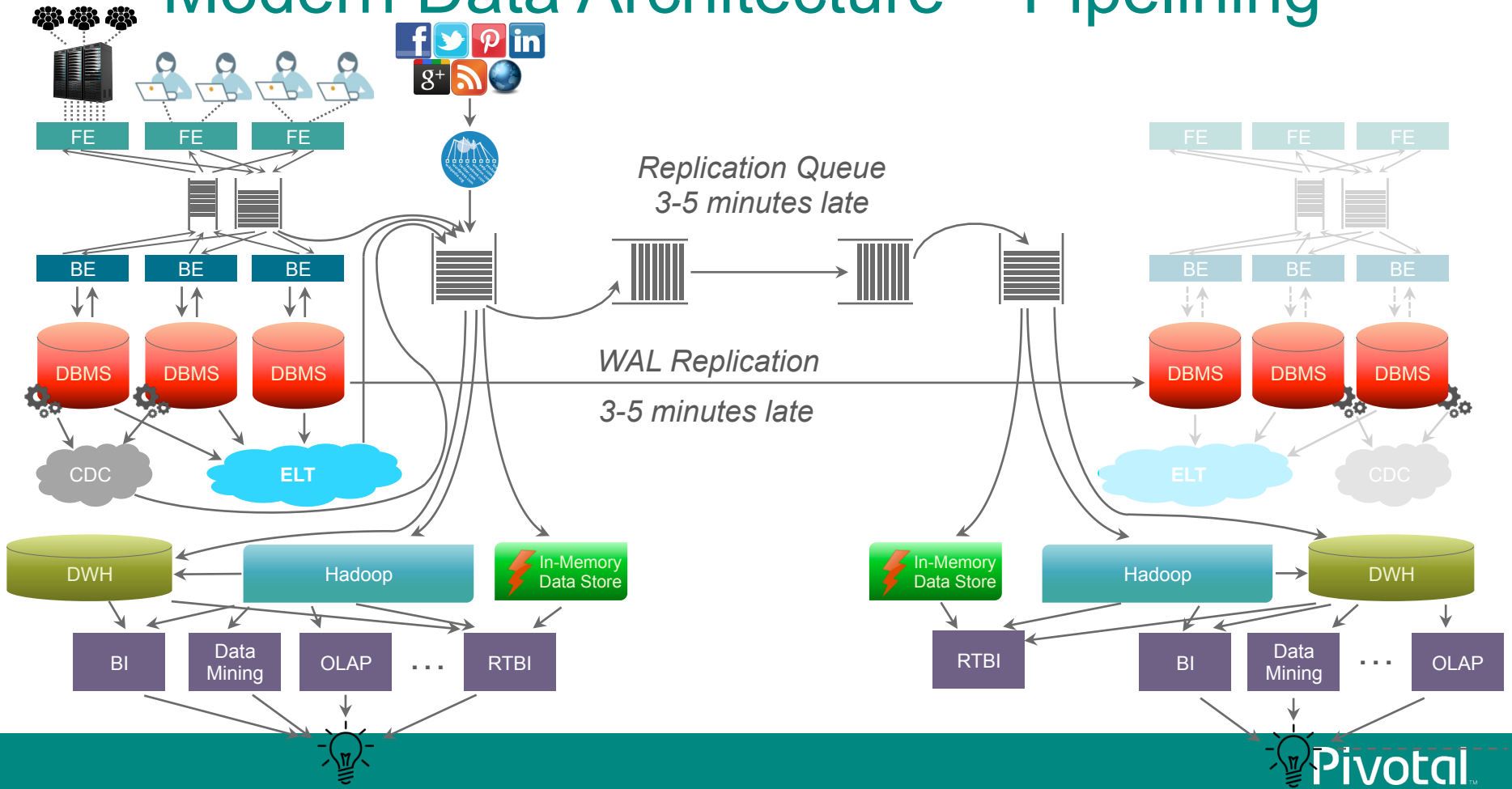
Modern Data Architecture – Pipelining



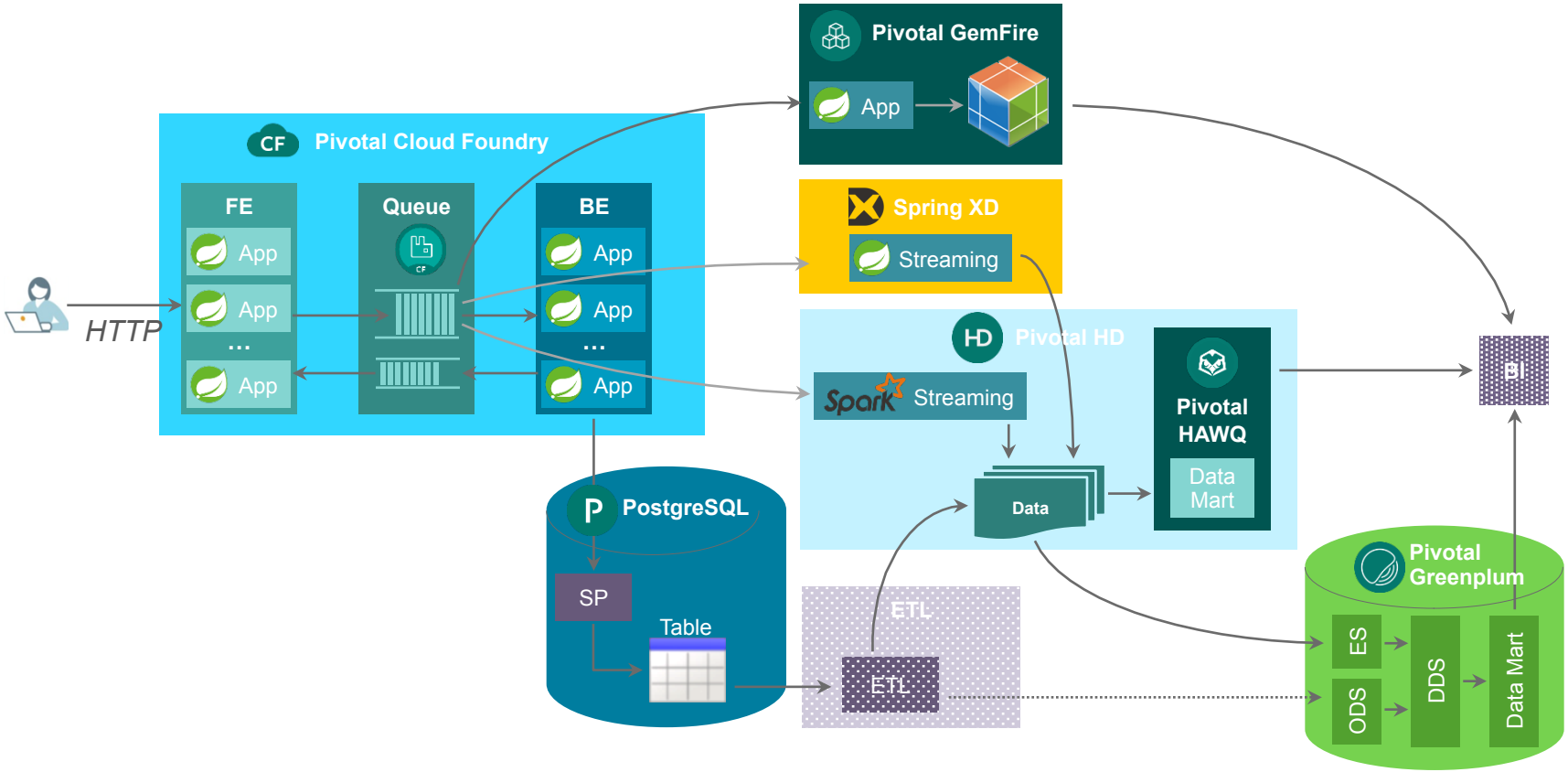
Modern Data Architecture – Pipelining



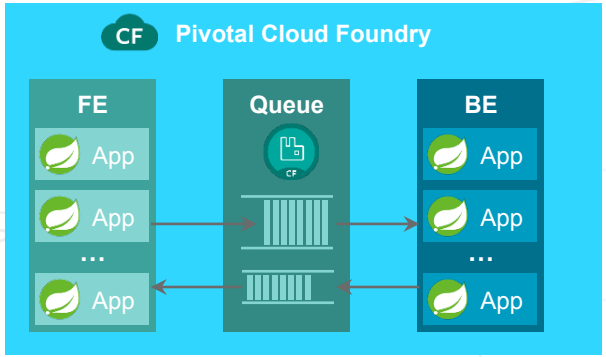
Modern Data Architecture – Pipelining



Pivotal and Modern Data Architecture

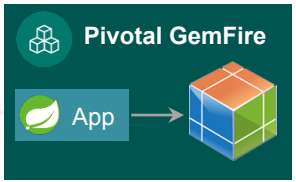


Pivotal and Modern Data Architecture



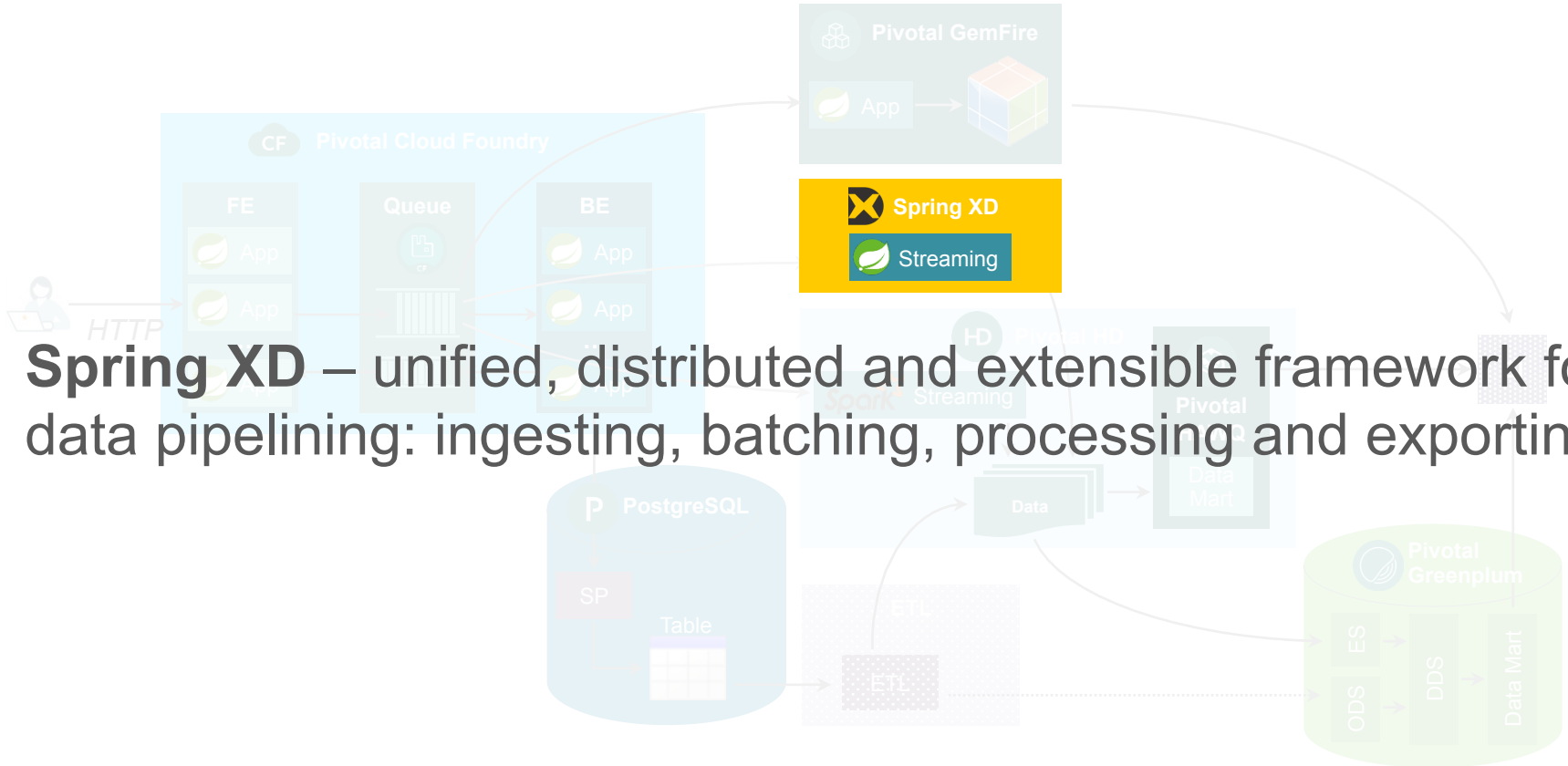
- **Pivotal Labs** – agile software development for next-generation applications
- **Pivotal Cloud Foundry** – PaaS for customer applications
- **RabbitMQ** – distributed message queue service on top of PCF
- **Spring IO** – foundation platform for modern applications

Pivotal and Modern Data Architecture



Pivotal GemFire and **Apache Geode (incubating)** – in-memory data grid enabling real-time data processing and real-time decision making for enterprises

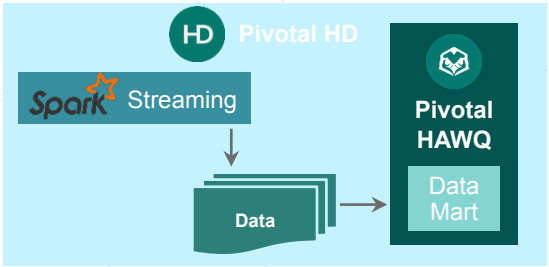
Pivotal and Modern Data Architecture



Spring XD – unified, distributed and extensible framework for data pipelining: ingesting, batching, processing and exporting

Pivotal and Modern Data Architecture

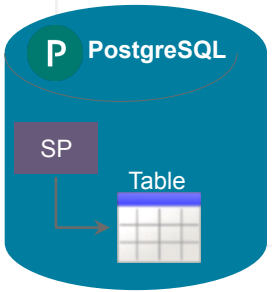
- **Pivotal HD** – leading Hadoop distribution based on ODP
- **Pivotal HAWQ** and **Apache HAWQ** (incubating) – bringing the power of MPP to the Hadoop cluster, best in class SQL-on-Hadoop solution



- **Apache Spark** – component of the Pivotal HD distribution, modern framework for distributed data processing

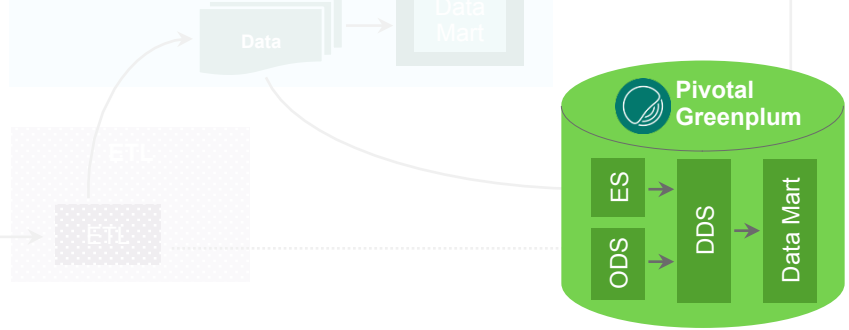
Pivotal and Modern Data Architecture

- **Pivotal PostgreSQL** – commercially supported by Pivotal open source distribution of PostgreSQL

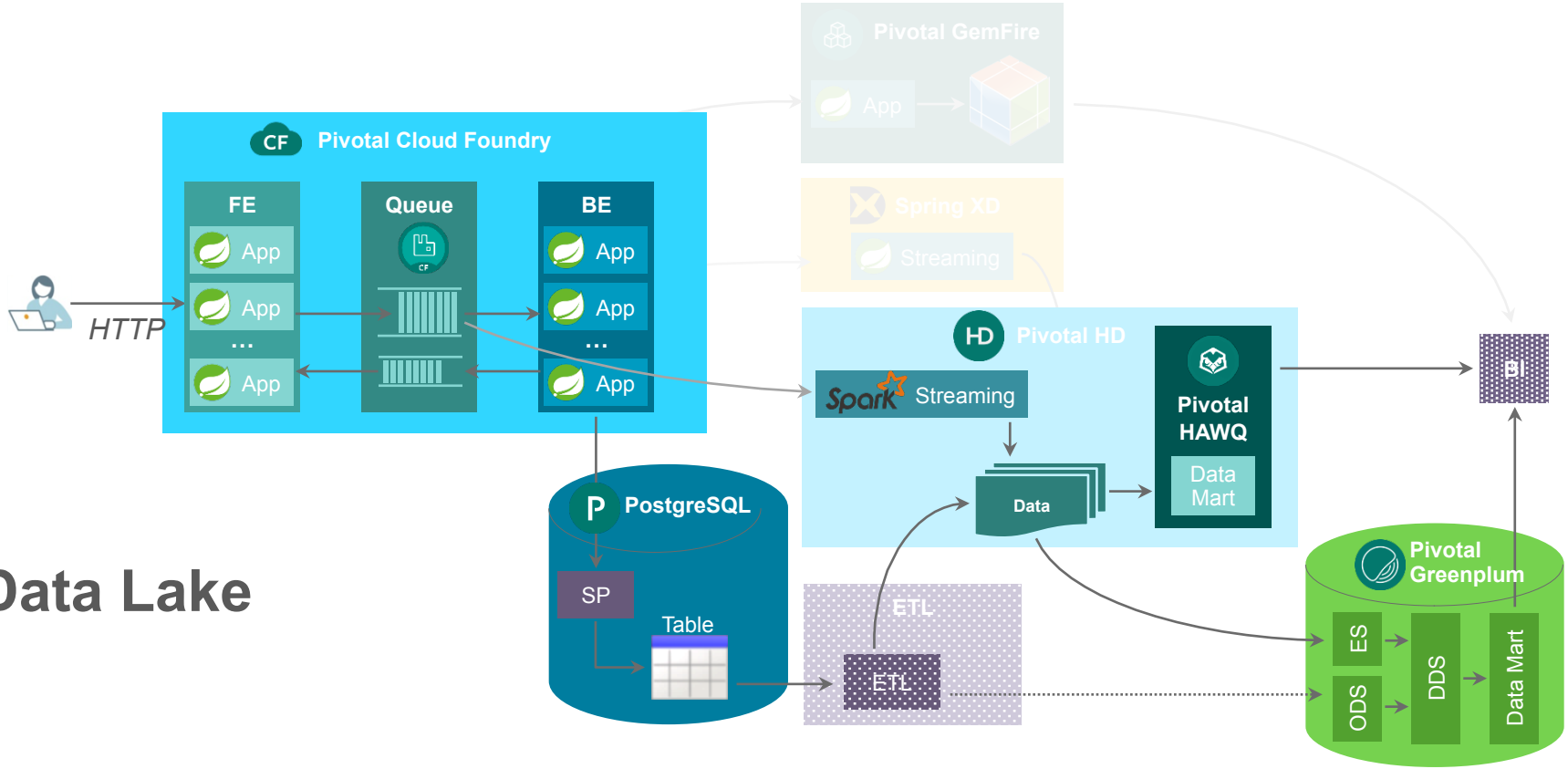


Pivotal and Modern Data Architecture

Pivotal Greenplum – leading analytical MPP database, foundation for the enterprise data warehousing systems and advanced analytics

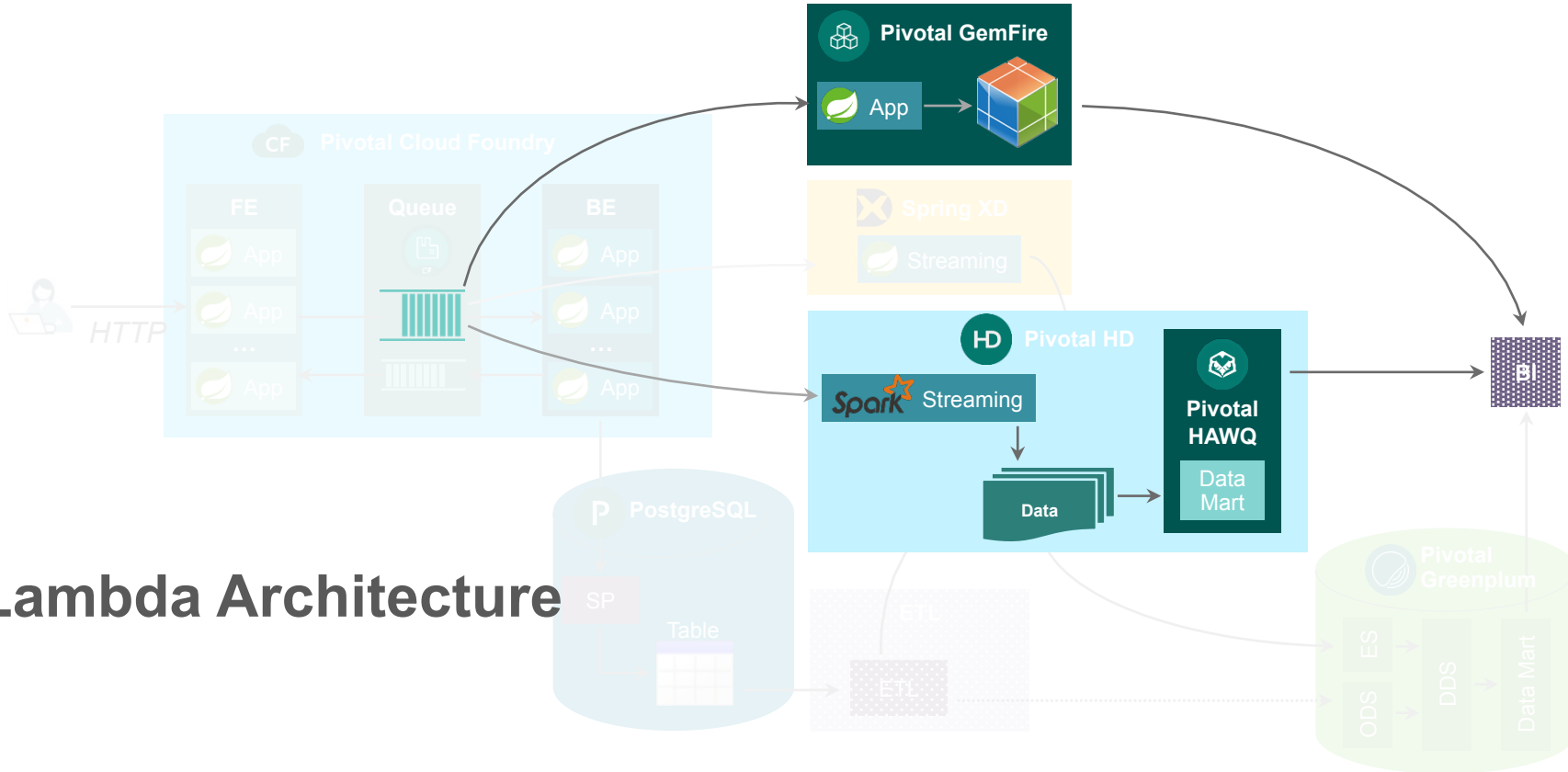


Pivotal and Modern Data Architecture



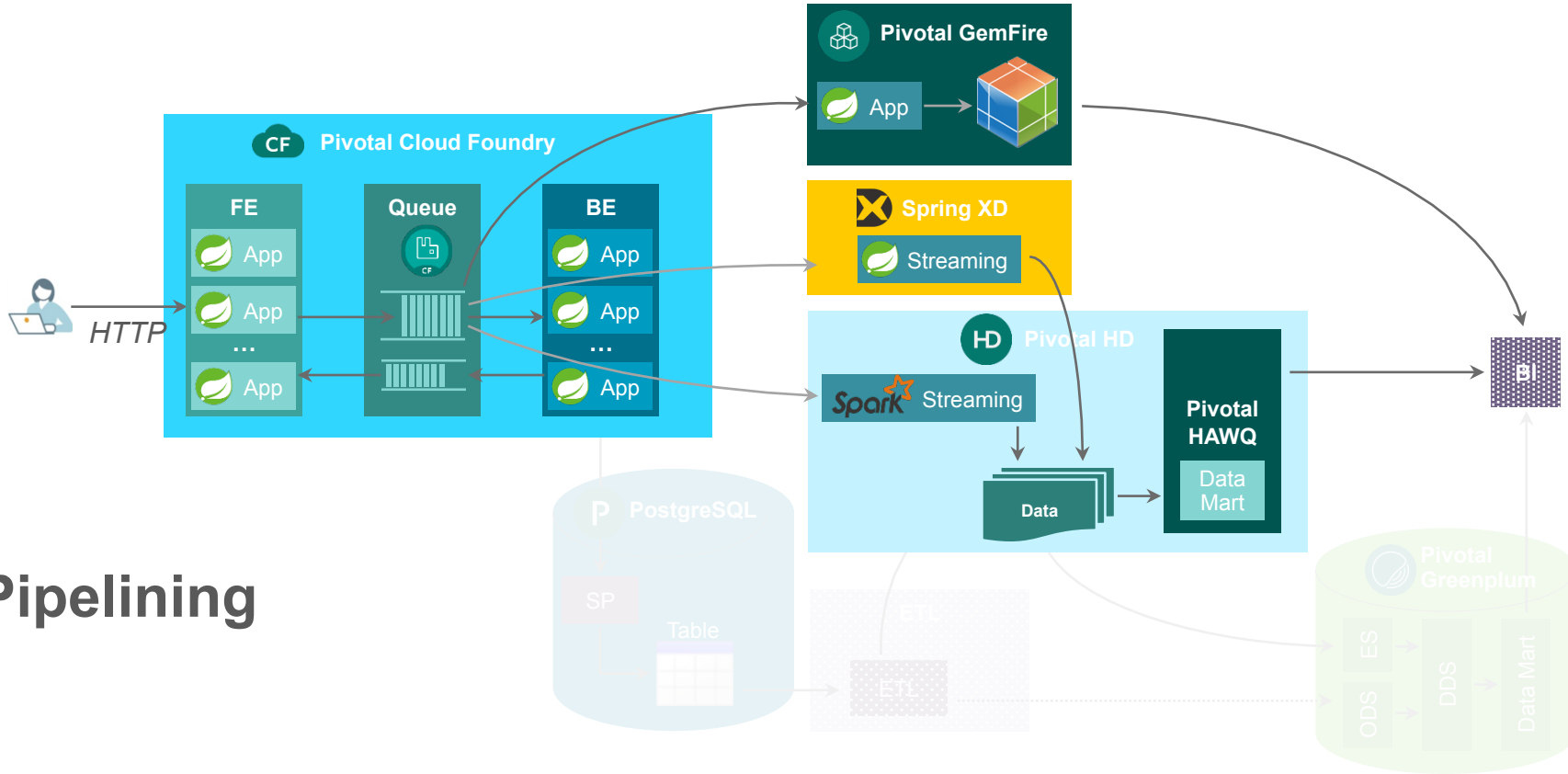
Data Lake

Pivotal and Modern Data Architecture



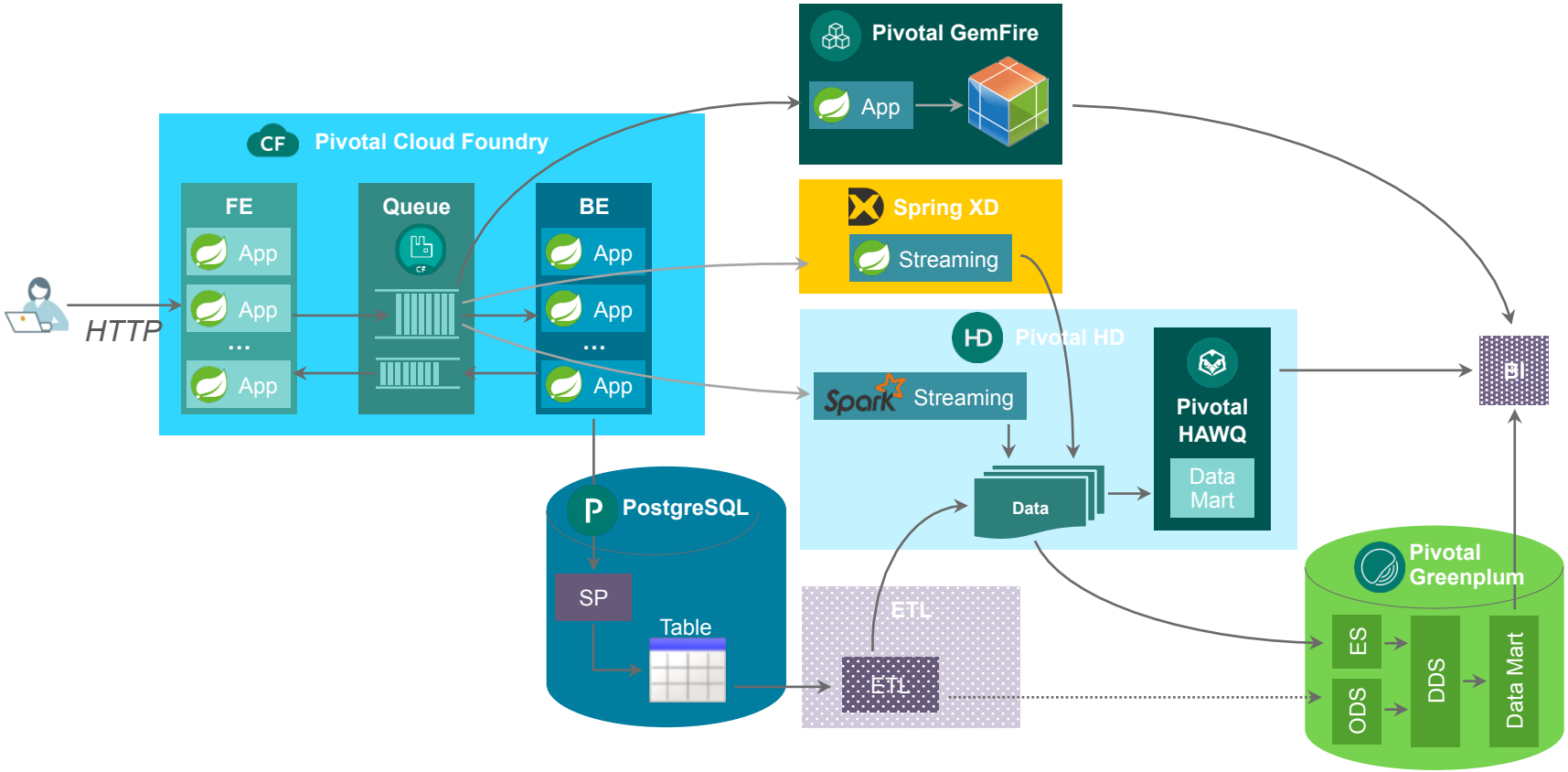
Lambda Architecture

Pivotal and Modern Data Architecture



Pipelining

Pivotal and Modern Data Architecture



Questions?

BUILT FOR THE SPEED OF BUSINESS