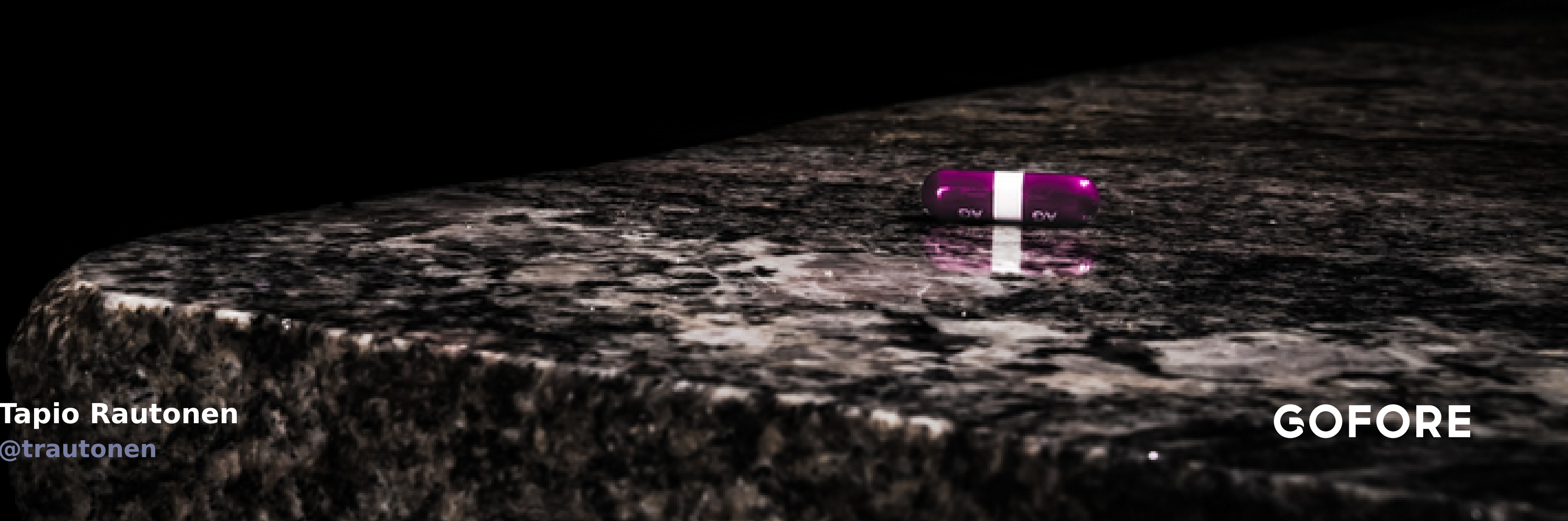
MAKING SENSE OUT OF YOUR BIG DATA



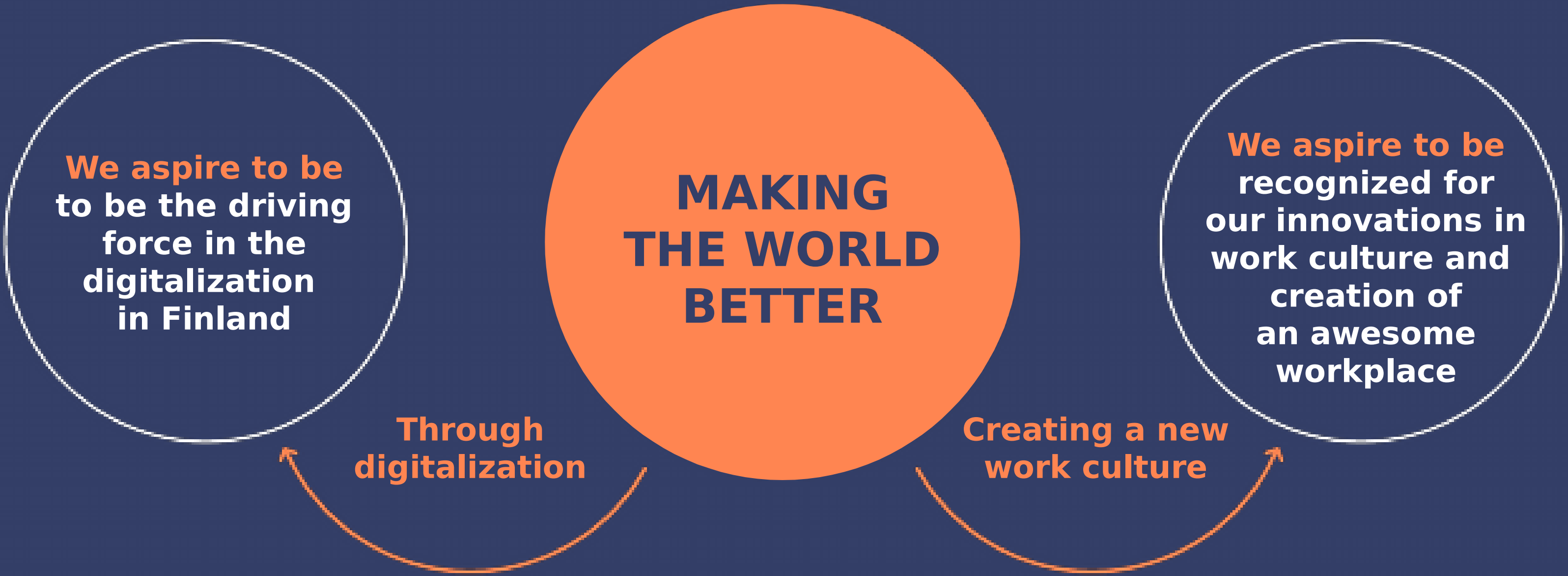
Tapio Rautonen
@trautonen

GOFORE

@trautonen

github.com/trautonen

fi.linkedin.com/in/trautonen



We aspire to be
to be the driving
force in the
digitalization
in Finland

MAKING
THE WORLD
BETTER

We aspire to be
recognized for
our innovations in
work culture and
creation of
an awesome
workplace

Through
digitalization

Creating a new
work culture

100%

I look forward to coming to work*

100%

Here we work as a team*

100%

Its fun to work here*

GREAT PLACE TO WORK®
Best
Workplaces 2016
Finland

* Trust Index 2015, Great Place to Work

how much is
BIG DATA

Tens of gigabytes

- Normal operational database
- Fits easily in a single machine
- Thousands of transactions per minute

Hundreds of gigabytes

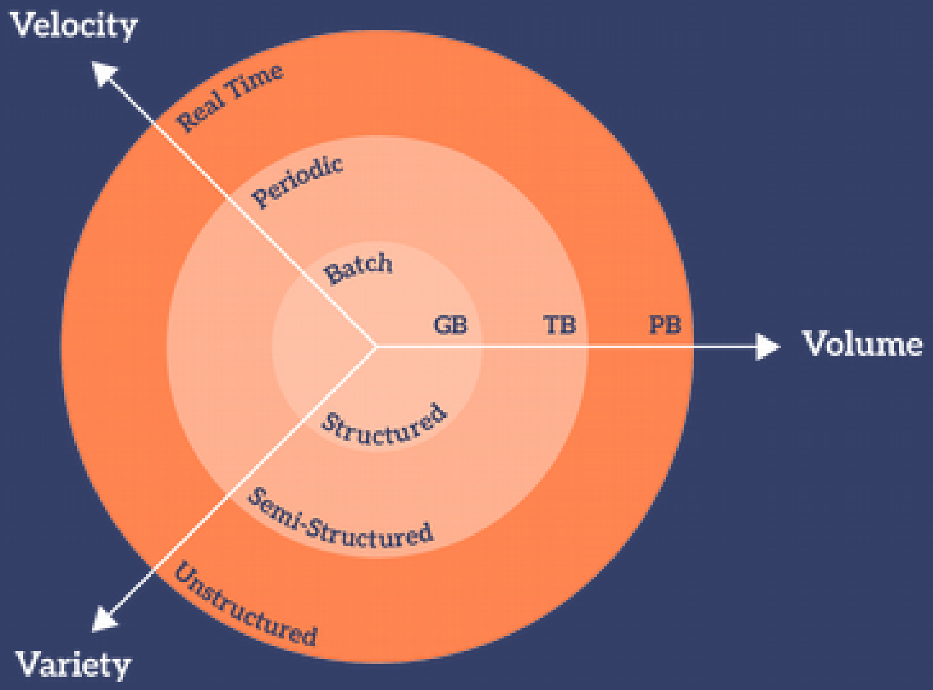
- Volume that global startups are dealing with
- Still reasonably priced hardware
- Traditional databases are capable of handling

“More data you can comfortably store and process in a single machine with traditional tools.”

DATA DISTRIBUTION

PARALLEL PROCESSING

VOLUME
VARIETY
VELOCITY



3-D Data Management: Controlling Data Volume, Velocity and Variety published in 2001 by Gartner

Tapio Rautonen
@trautonen

GOFORE

DATA WAREHOUSE

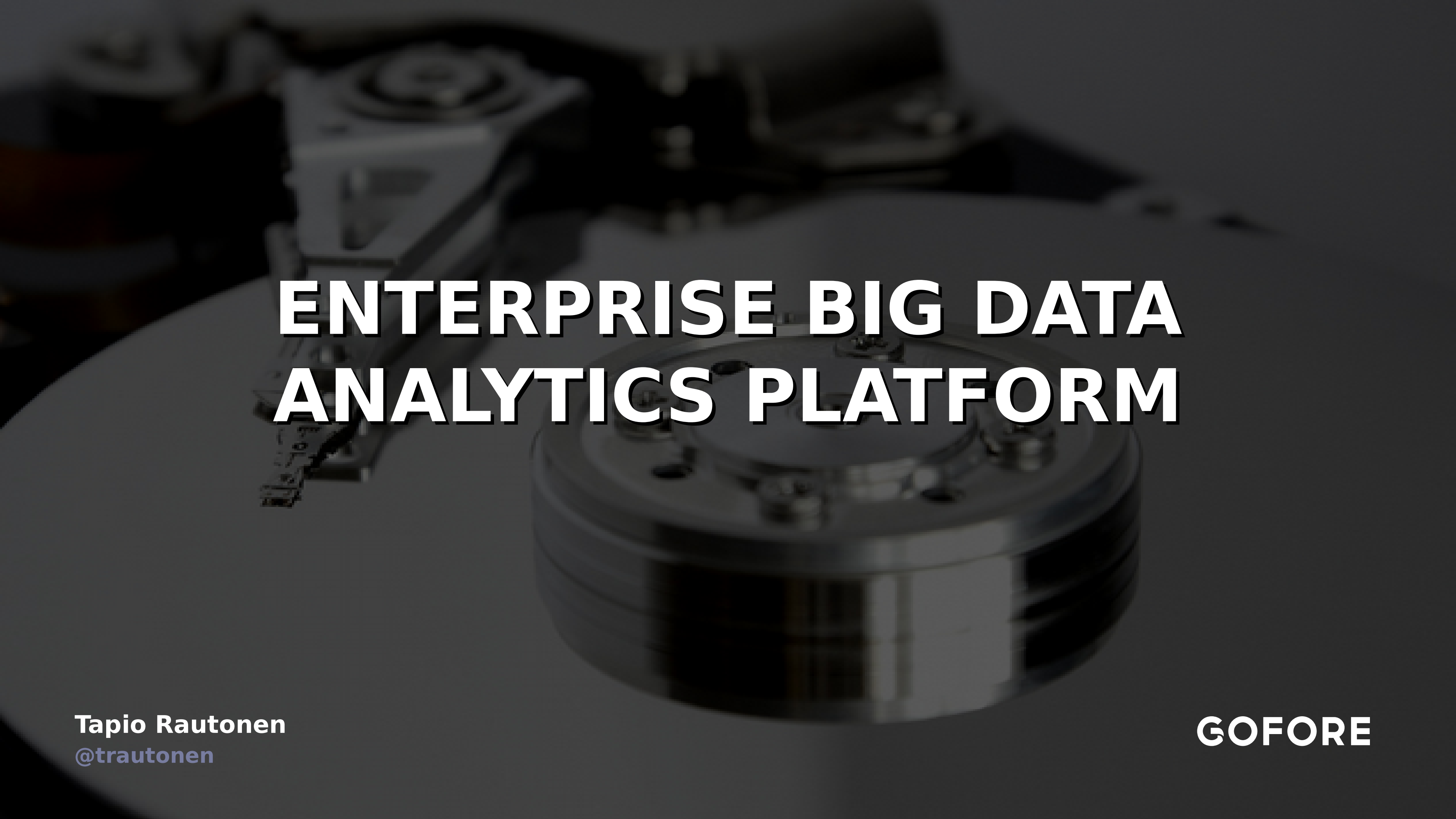
- structured
- schema-on-write
- only modeled data is stored
- expensive to store huge amounts of data
- cheap and fast to process to some extent
- good security models
- easy to integrate

DATA LAKE

- raw, unstructured
- schema-on-read
- everything can be stored
- cheap to store huge amounts of data
- expensive and slow to process
- unmature security models
- complex integrations

“But which camp should I choose?”

the only winners are the consultants



ENTERPRISE BIG DATA ANALYTICS PLATFORM

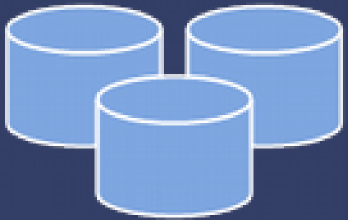
Tapio Rautonen
@trautonen

GOFORE

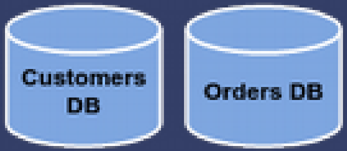
Data distribution

- How to control distribution and scaling?
- How to process data when you cannot access everything?
- How to identify your data from various sources?
- How to query efficiently from distributed data store?

Clustering



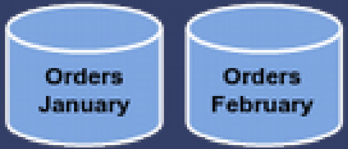
Database Federation



Customers DB

Orders DB

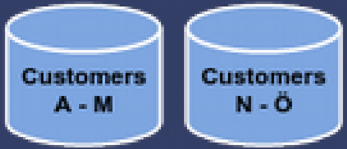
Table Partitioning



Orders January

Orders February

Table Sharding



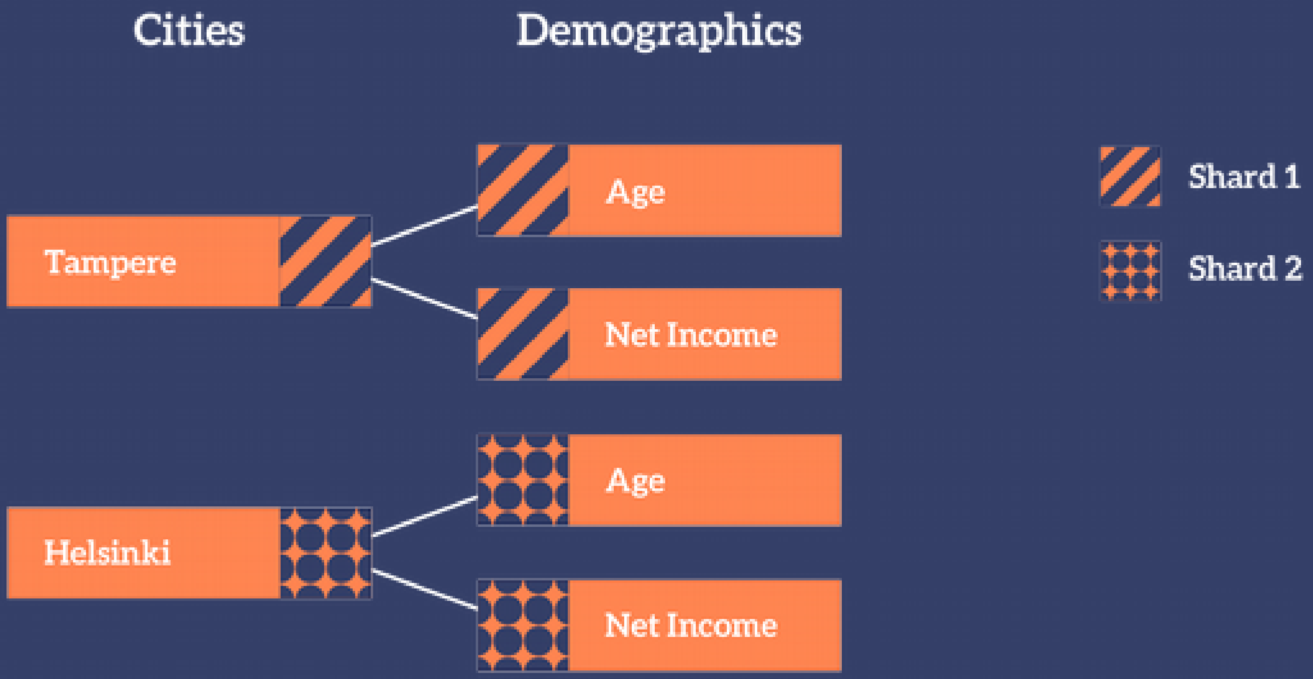
Customers A - M

Customers N - Ö

KEYS
distribute & identify

Distribution (or sharding) key

- Split data to multiple storage locations based on distribution key.
- Columnar storages are a lot more effective for analytical queries than row based storages.
- Routing requires some overhead and rebalancing of shards is really expensive.


Cities
Demographics
Tampere
Helsinki
Age
Net Income
Age
Net Income
Shard 1
Shard 2

“Dimension table's primary key and fact table's corresponding foreign key should be the distribution keys.”

Identification keys

- Natural key
  key formed of attributes that already exist in real world

- Business key
  key formed of attributes that already exist in business systems

- Surrogate key
  generated key with no business meaning

To hash or not to hash

- Sequences are bottlenecks due to dependencies and global state
- Hashes are easy to represent as ASCII text and transfer between different storage systems
- Hashing can combine compound keys, but are vulnerable to collisions
- Hashes require more storage and index space and might affect distribution

PROCESS
divide & conquer

Tapio Rautonen
@trautonen

GOFORE

Parallel processing

- MapReduce is not a silverbullet
- Only small portion of the data fits in memory
- Problems come in different forms: streams, graphs, documents
- Fallacies of distributed computing by Peter Deutsch

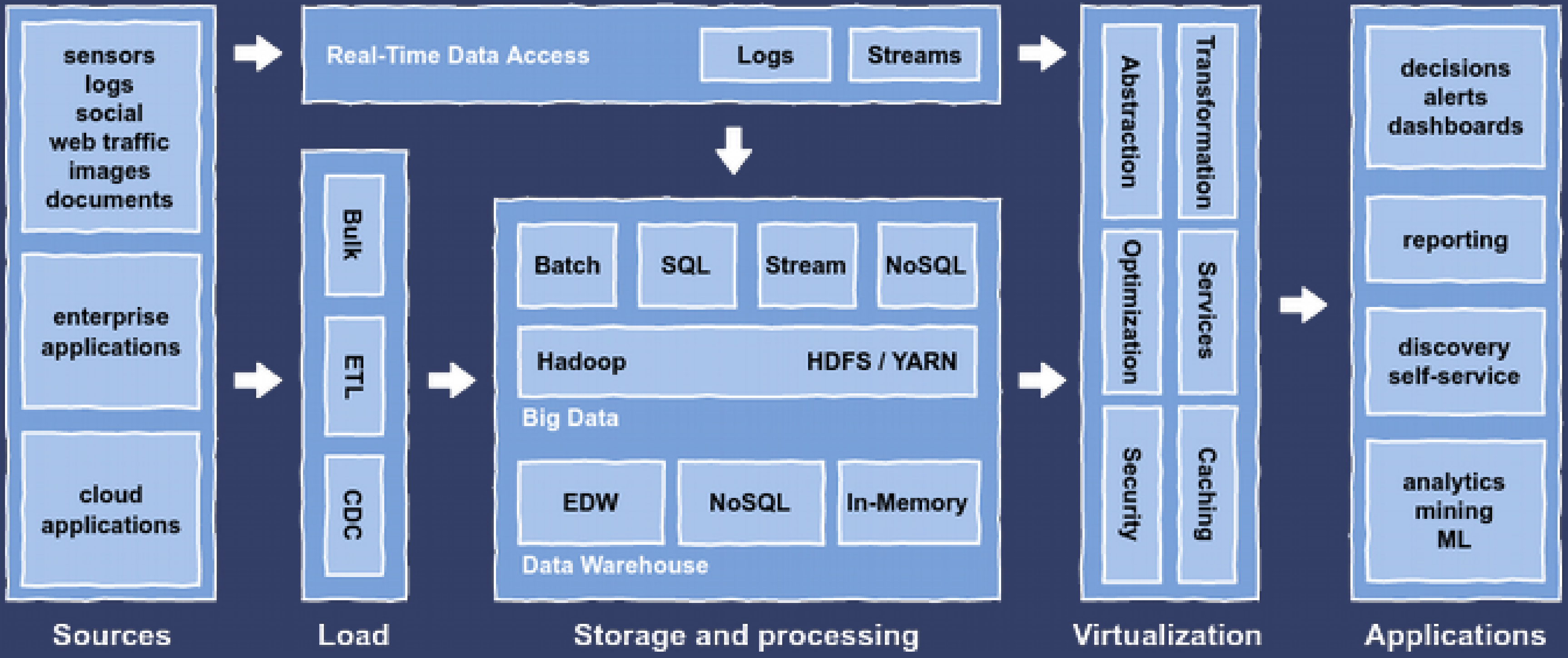
Scaling

- I/O, memory or CPU bound?

- Changing data distribution is expensive

- Vertical scaling increases the capacity of processing nodes

- Horizontal scaling increases the number of nodes (parallelism)



http://www.datavirtualizationblog.com/logical-architectures-big-data-analytics/

Data loading

- Sqoop
  bulk transfers of data between Hadoop and structured datastores

- ETL (extract, transform, load)
  different forms like ESB, lambda, microservice or reactive stream

- CDC (change data capture)
  determine and track data changes to react when data is changed

Stream and event inputs

- IoT devices, logs, events
  small payloads, huge volume and velocity

- Collect and batch
  target systems handle larger batches more efficiently

- Hosted and on-premises solutions
  Flume, Kafka, AWS Kinesis/Firehose, Google Cloud Pub/Sub

Hadoop ecosystem

- HDFS for distributed storage and YARN for resource management
- Batch processing (MapReduce, Tez)
- Data warehouse and SQL (Hive, Spark, Drill)
- Stream (Spark, Flink)
- NoSQL (HBase)

Data warehouse

- Modern data warehouse is not just RDBMS
  combination of RDBMS, NoSQL and In-Memory databases

- Cloud databases as a service
  AWS Redshift, Google BigQuery

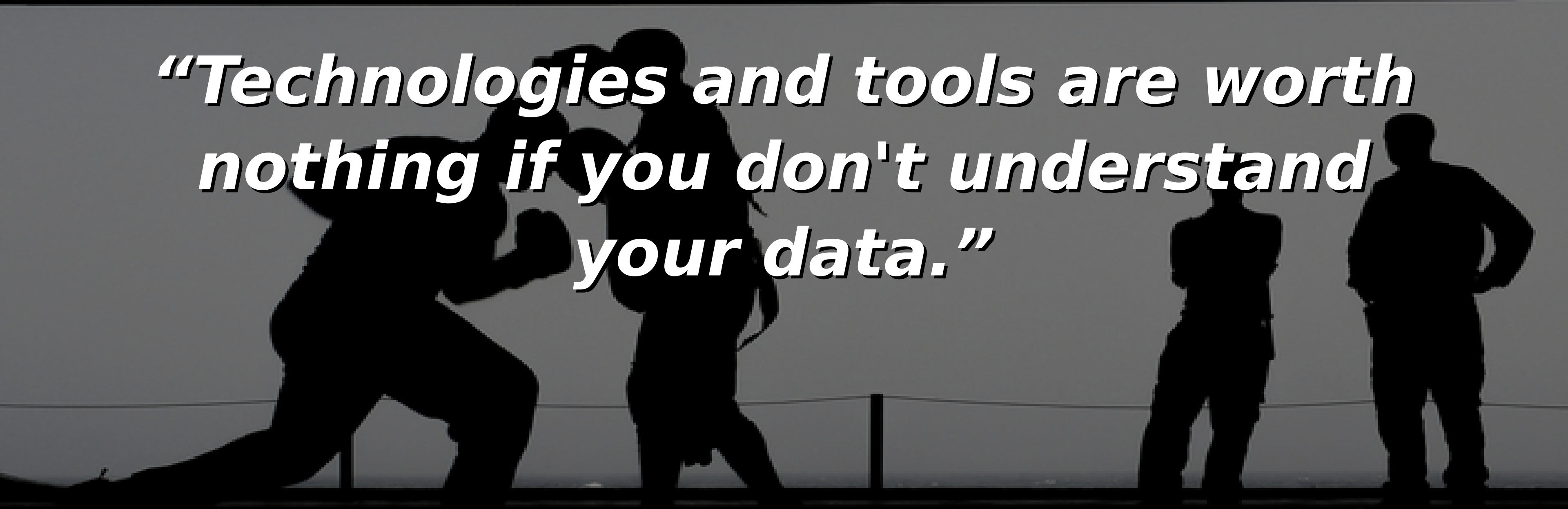
- Data Vault 2.0
  not just technology, also methodologies for project management

Data virtualization

- Data virtualization provides information agility
  combines data warehouse, big data and other data sources

- Late binding to many unresolved issues
  abstraction, transformation, optimization and security

- Unified data access services to all clients
  data sources accessible in different format with access control



“Technologies and tools are worth nothing if you don't understand your data.”

embrace master data management

lean and agile development

pick the right tools

THE END

Tapio Rautonen
@trautonen

GOFORE