SJK006 - Master in Intelligent Systems

Big Data Architectures

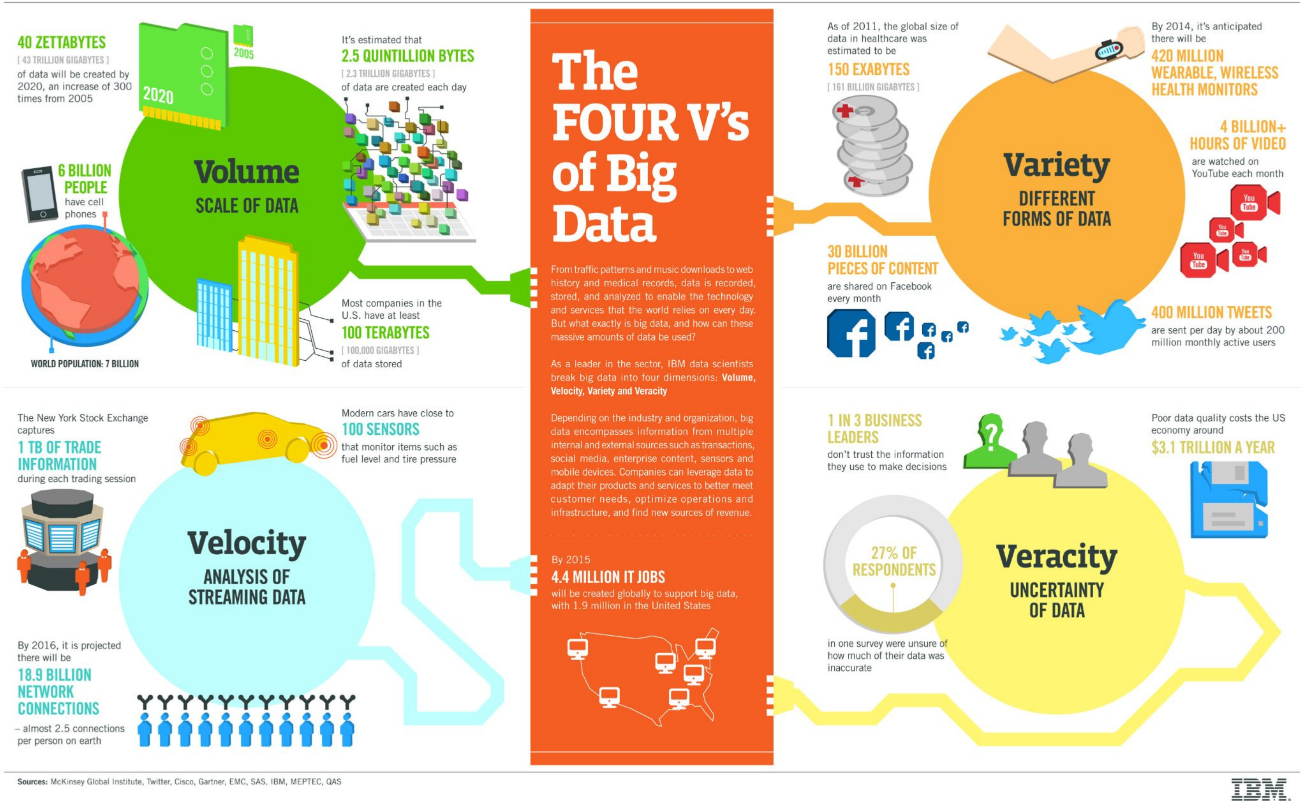
Implementing Data Value Chains at large scale

Summary

We are going to

- recap. Big Data concepts
- introduce a Use Case for BD architectures
- analyze the logic of BD architectures
- introduce the main ETLing concepts
- practice with a Use Case example

The 4 V's of Big Data



Big Data Value Chain

Data Acquisition

- Structured data
- Unstructured data
- Event processing
- Sensor networks
- Protocols
- Real-time
- Data streams
- Multimodality

Data Analysis

- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked Data
- Data discovery
- 'Whole world' semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data Curation

- Data Quality
- Trust / Provenance
- Annotation
- Data validation
- Human-Data Interaction
- Top-down/Bottom-up
- Community / Crowd
- Human Computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data Storage

- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and Performance
- Data Models
- Consistency, Availability, Partition-tolerance
- Security and Privacy
- Standardization

Data Usage

- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualisation
- Modeling
- Control
- Domain-specific usage

Use Case

Intelligent Transport

- Global view of the transport scenario in a city
- Gather data from all the perspectives
  - Sensor data
  - Citizen interactions
  - Transport Modalities
  - City's Infrastructures
  - etc.

Smart Cities for Public Transportation

TRANSPORTS

interactions

BIG DATA
(big picture)

CITIZENS

Data Chain Values

DATA SOURCES

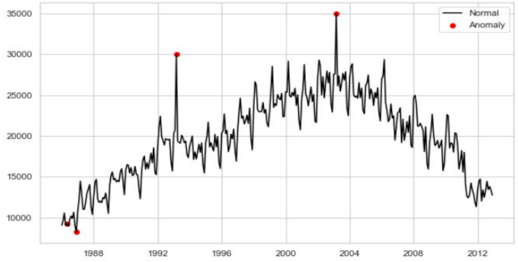
- BiciCas
- City Bus
- Geo Loc
- Shops
- Social networks
- Open Street Maps
- Open Data

DATA Processes

- Alerts
- Capacity
- Demand
- Mobility
- Businesses
- Incidents (plans)

Visualizations and insights

Real-time alerts and reporting



Rent-a-bike datasets (PayNoPain)

benches
bench_status
benches_anchors

benches_messages
benches_notifications

anchors

incidents

loans

putbacks

bicycles

bans
users
subscriptions
payments

Datasets/tables

anchors(uuid, bench_uuid, number)
bans(uuid, use_uuid, date_start, date_expire, date_created, reason)
benches_anchors(bench_uuid, anchor_uuid, anchor_number, bicycle_number, date_created)
benches(uuid, name, latitude, longitude, date_created)
benches_notifications(uuid, bench_uuid, notification, priority, date_created)
bench_messages(uuid, message, date_start, date_expire, date_created)
bicycles(uuid, number, date_created, status)
incidents(uuid, error_key, use_uuid, anchor_uuid, status, date_created, bike_uuid)
loans(uuid, use_uuid, bicycle_uuid, anchor_uuid, date_created, type_access)
putbacks(uuid, bicycle_uuid, anchor_uuid, date_created)
users(change_uuid, uuid, ~~document~~, ~~name~~, ~~surname~~, ~~street~~, city, postcode, birthday, ~~phone~~, ~~email~~, gender, date_created, date_change, deleted)

bench_status(bench_uuid, date_lastseen, ip, queue, version, number_loans)
loan_historical(uuid, loan_uuid, use_uuid, bicycle_uuid, loan_anchor_uuid, loan_date_created, putback_uuid, putback_anchor_uuid, putback_date_created, type_access, app_loan)

Exercise: data chain values

HISTORICAL

Incidentes por demografía, geografía

Rutas más frecuentadas + temporalidad

Histórico de ocupación de bancadas con respecto a la demanda

Simulación de movimiento de recursos (modelo generativo probabilistico)

Estudiar el “estado” de calidad de las bancadas y las bicicletas (kms que lleva la bicicleta)

Recorrido de una bicicleta.

Popularidad de las bicicletas.

Bicicletas con problemas técnicos (se devuelven en la misma bancada con un tiempo corto entre loan y putback).

Mapa por código postal de demanda.

STREAMING

Incidentes: cálculo real de bicicletas operativas

Situación actual + modelo predictivo de demanda -> movilidad de recursos

Avisos dirigidos

Modelo predictivo de posible “relleno” de bicicletas en bancadas vacías

Recomendación de bancadas cercanas y tiempo previsto (localización del usuario)

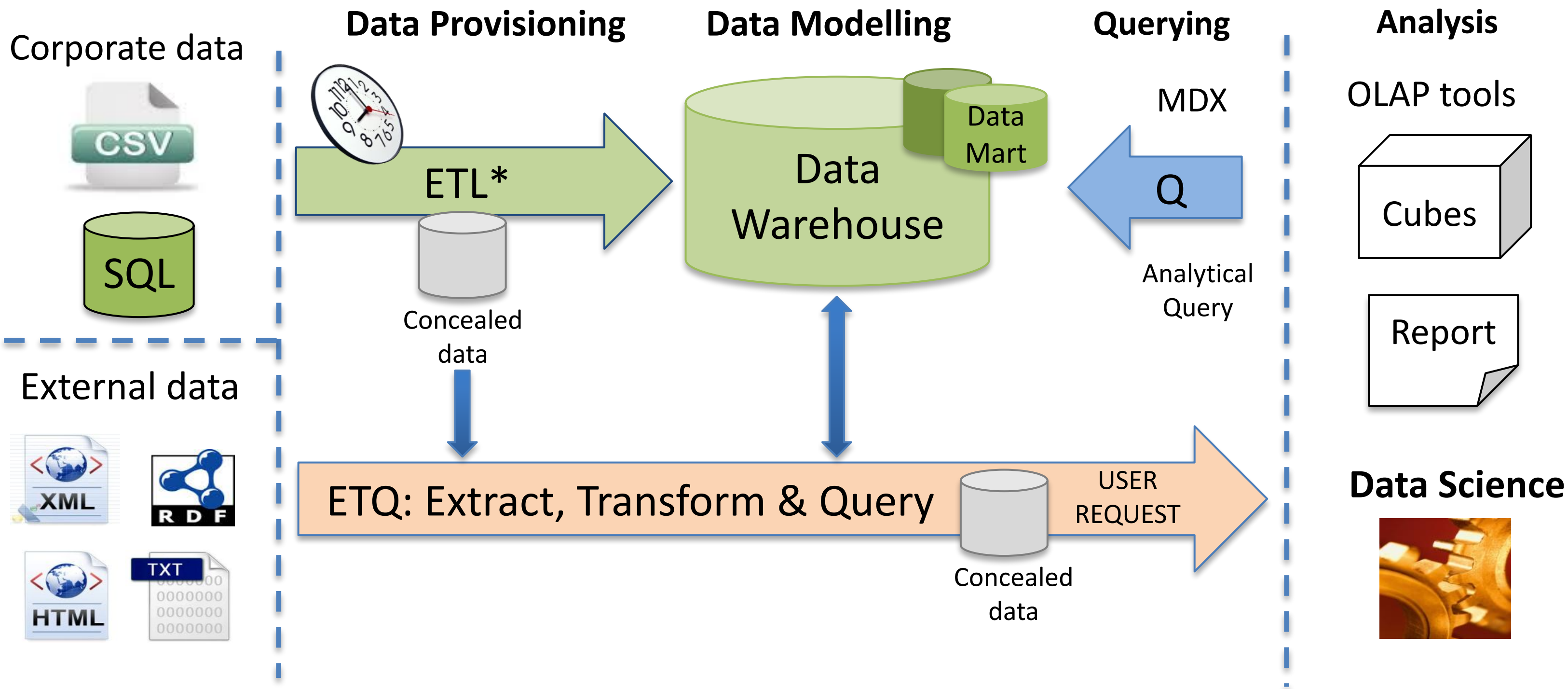
Alerta de bancada llena.

IMPLEMENTING DATA VALUE CHAINS

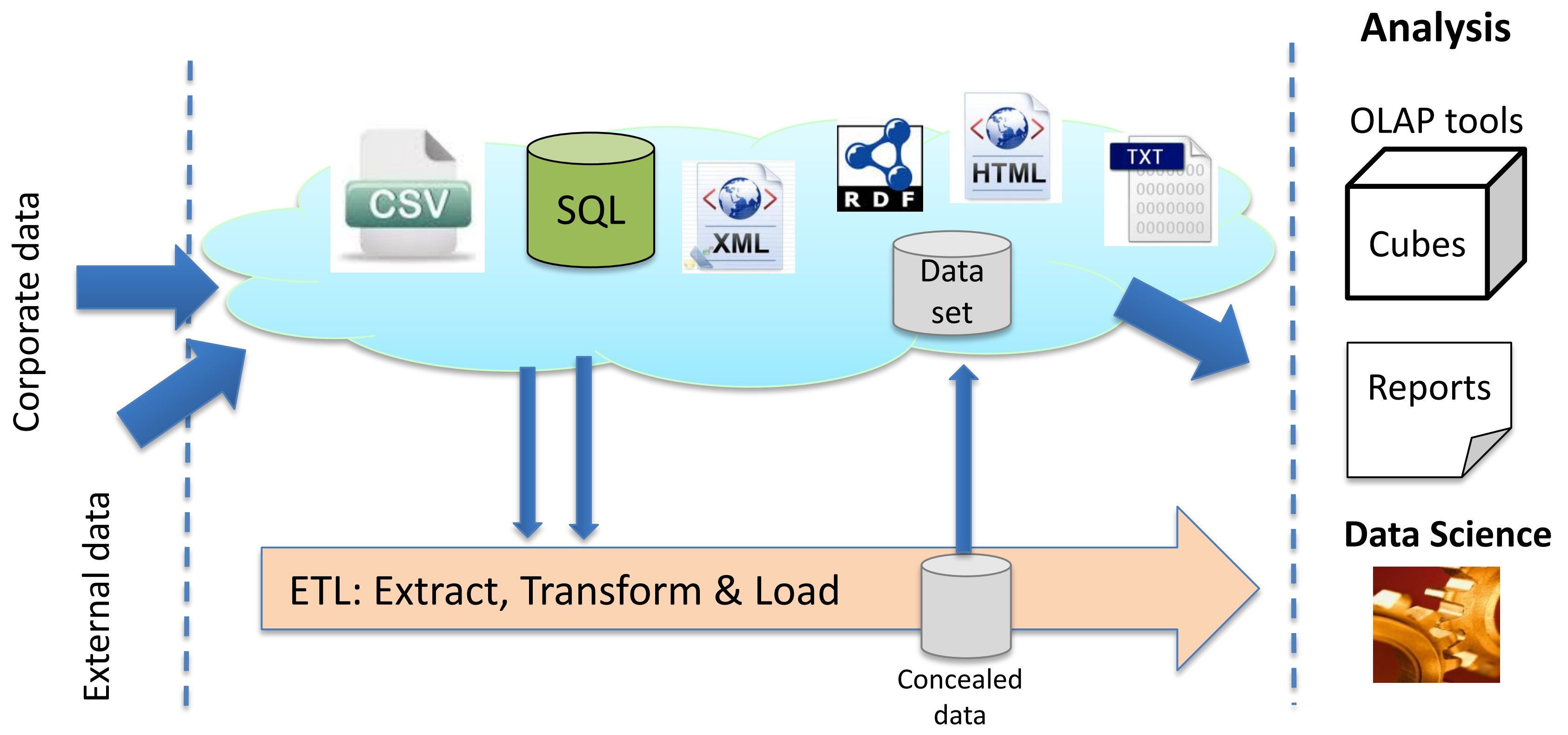
Big Data Architectures

- Data warehousing
- Data lakes
- Kappa architecture
- Lambda architecture

DATA WAREHOUSING



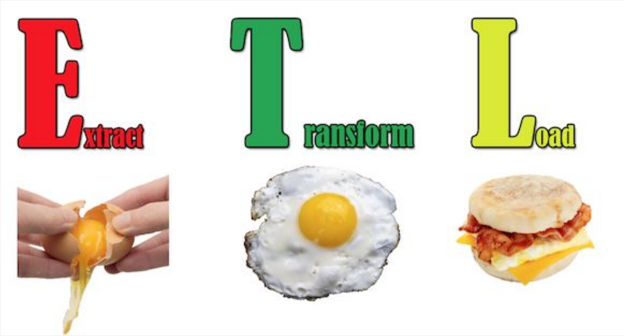
DATA LAKES



Corporate data
External data
CSV
SQL
XML
RDF
HTML
TXT
Data set
ETL: Extract, Transform & Load
Concealed data
Analysis
OLAP tools
Cubes
Reports
Data Science

ETL/Q

Basic operations for capturing and transforming data



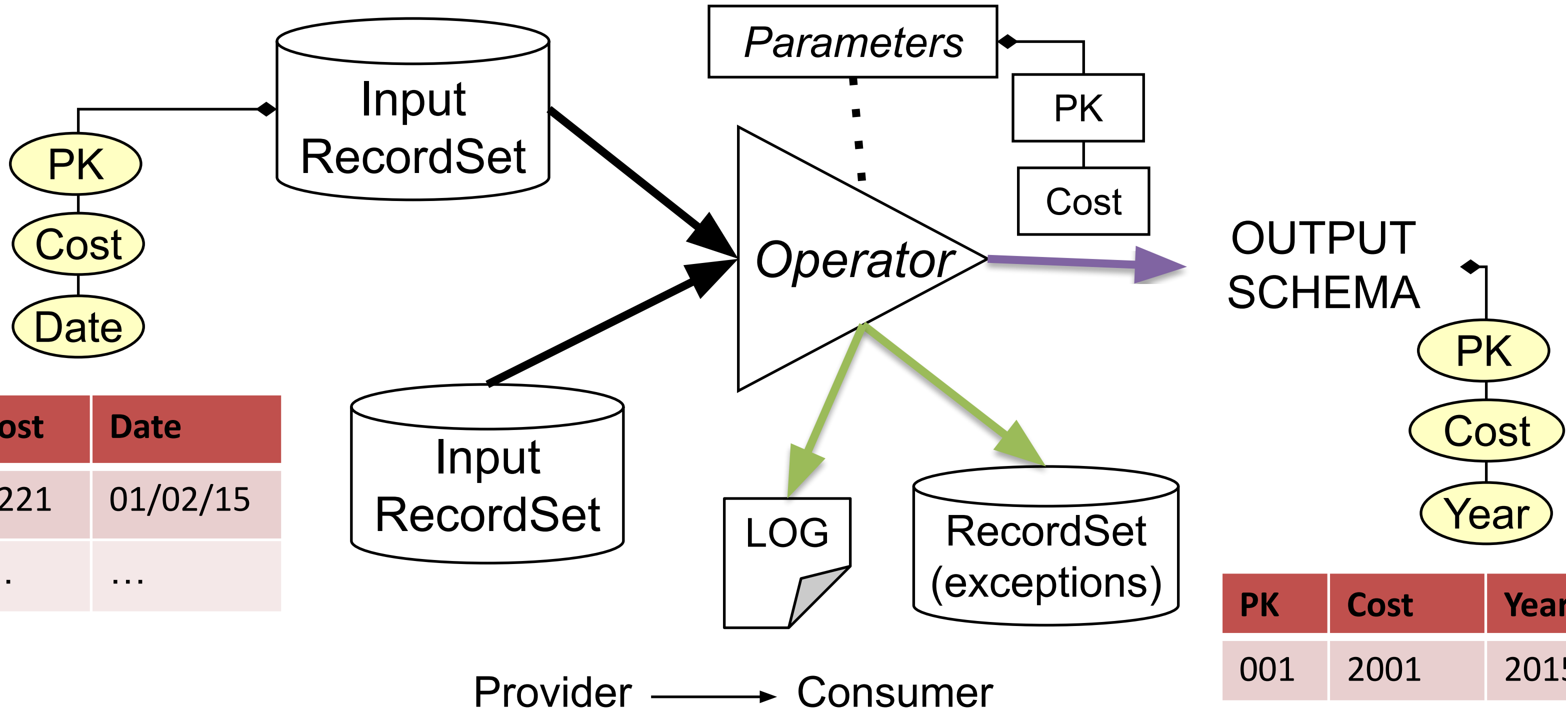
An ETL is a workflow of transformations on source data, which is aimed at **integrating** and **concealing** data for analytical purposes

Improving the quality of data

- **Data Cleaning**:
  - Detect and fix omitted or missing values
  - Remove noisy data (no relevant for analysis)
  - Detect incoherent data
- **Data integration**:
  - Duplicate removal
  - Gathering data of the same entities
  - Homogenization of terms and keys
- **Data transformation**:
  - Change of formats, scaling and normalization
- **Data reduction**:
  - Selection of attributes and aggregation of data

ETL: Workflows of transformation operators

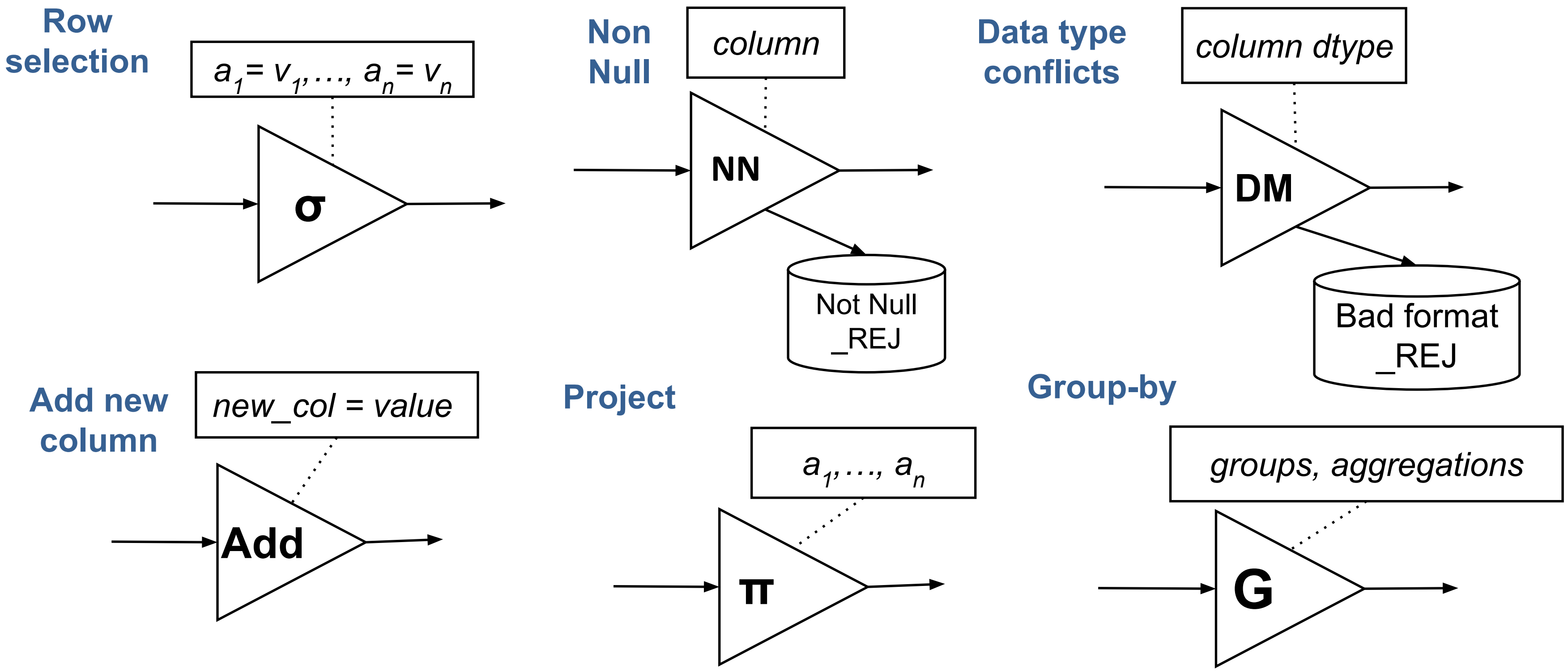
Vassiliadis et al. Information Systems 30 (2005)



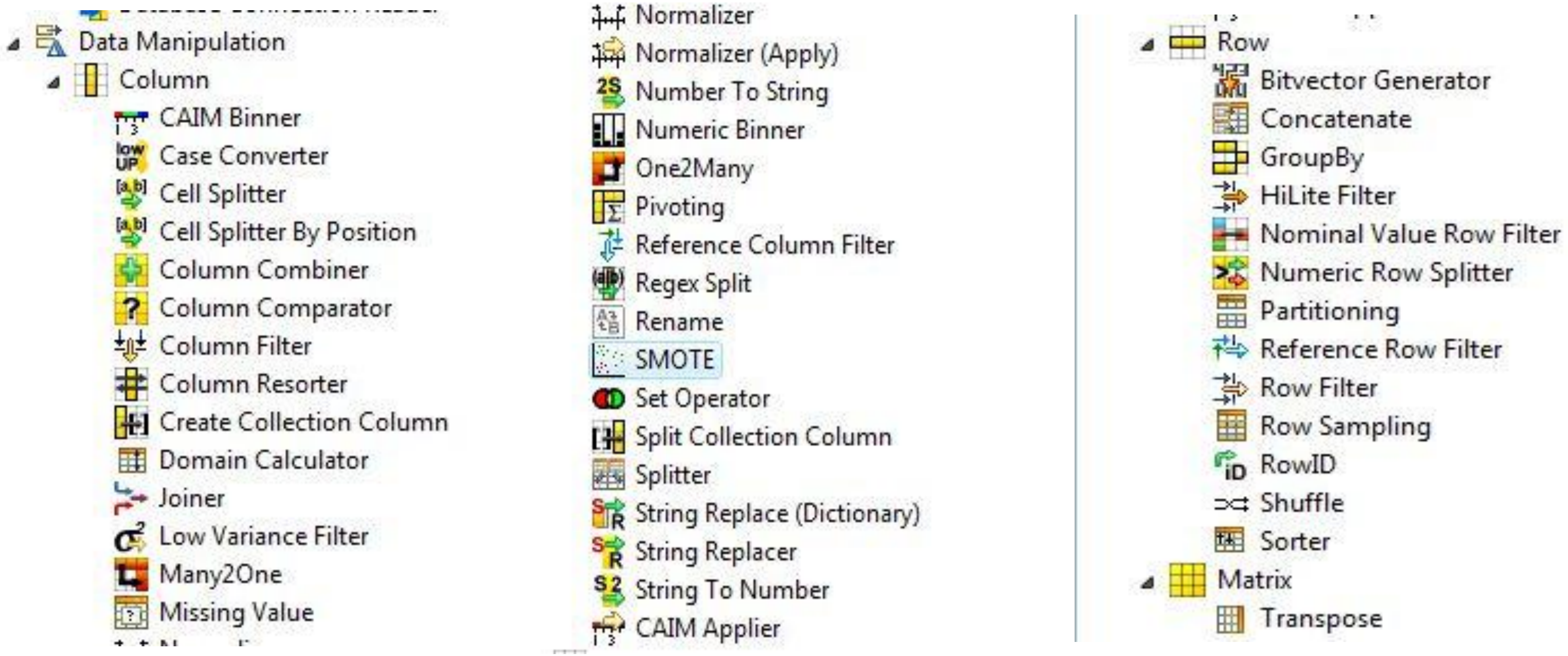
PK	Cost	Date
121	1221	01/02/15
…	…	…

PK	Cost	Year
001	2001	2015

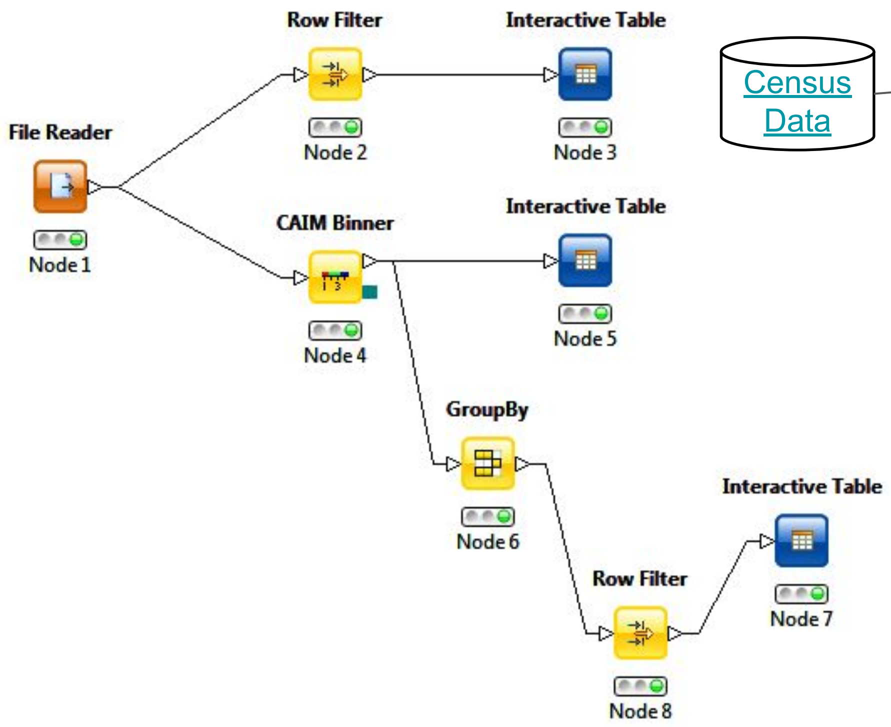
Some examples of operators

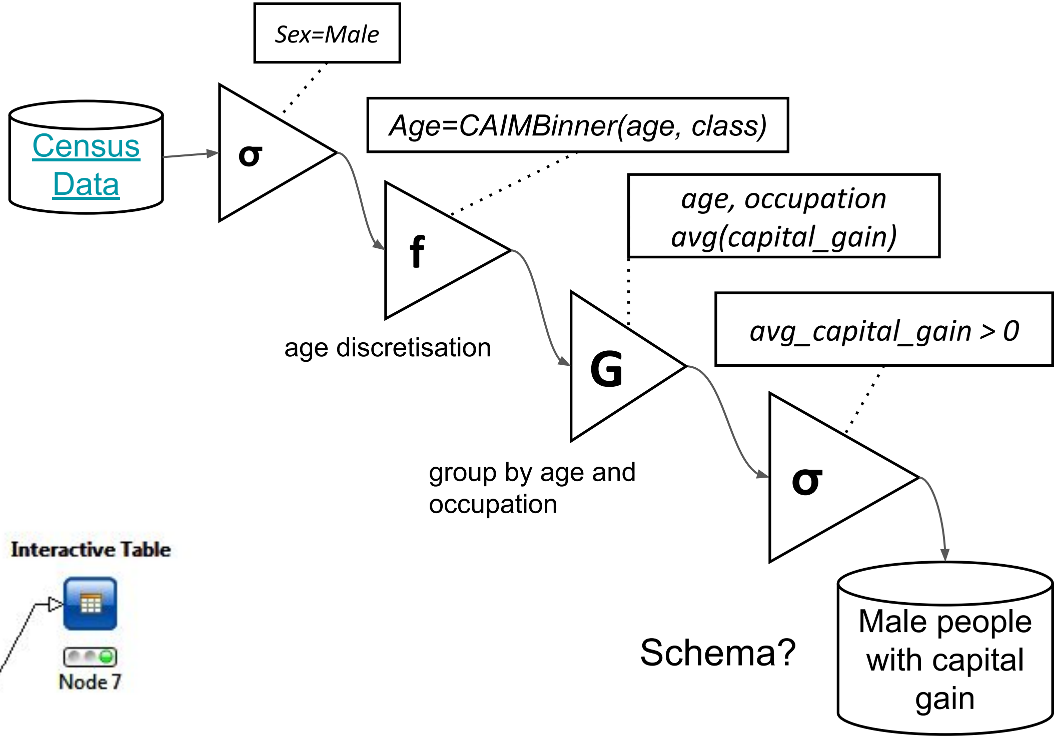


Tools for ETLing: KNIME



Tools for ETLing: KNIME



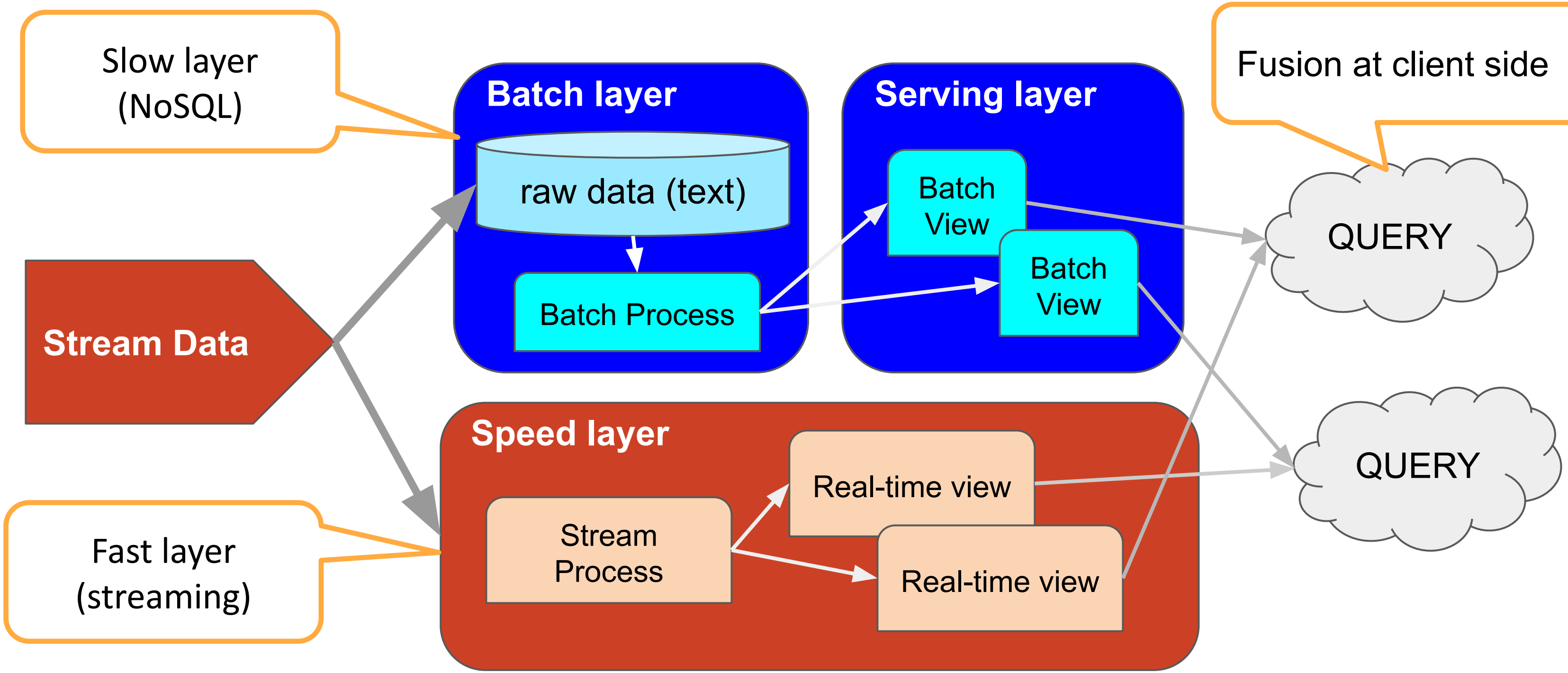


Exercise: Pandas for ETLing

SELECTION
DUPLICATE REMOVAL
MISSING VALUES
GROUP-BY
JOIN
DIFFERENCE (anti-join)
UNION of recordsets
PROJECT columns
CONCAT columns
SPLIT a column
APPLY a function to a column (or several)
DISTRIBUTION (reverse of join)

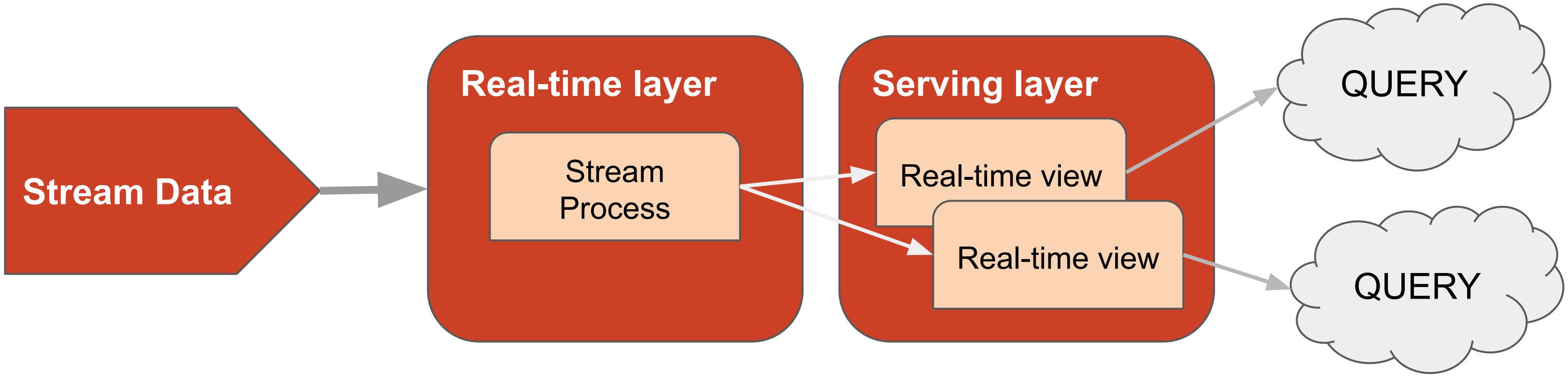
(complete with Pandas methods)

Lambda Architecture (Streaming data)

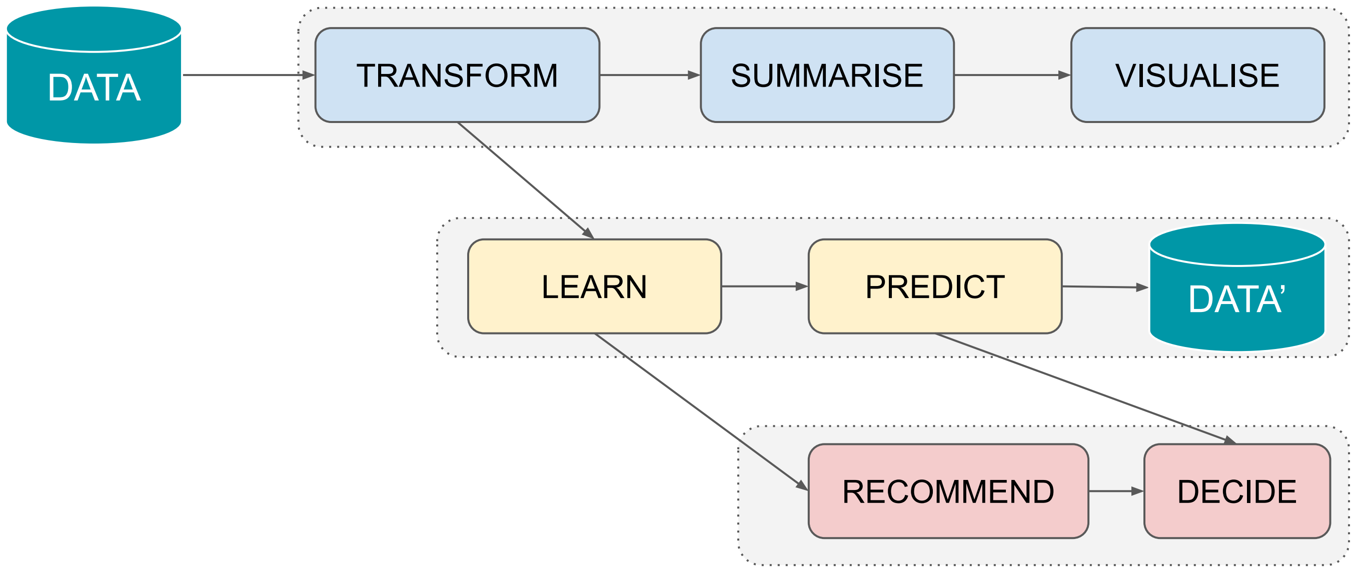


Kappa Architecture

- Use of “logs” (append-only files) to store data.
- The “batch” layer of the Lambda architecture is removed.
- Greatly simplifies code maintenance.
- Great processing speed.



Lambda or Kappa?



Descriptive Analysis

Predictive Analysis

Prescriptive Analysis