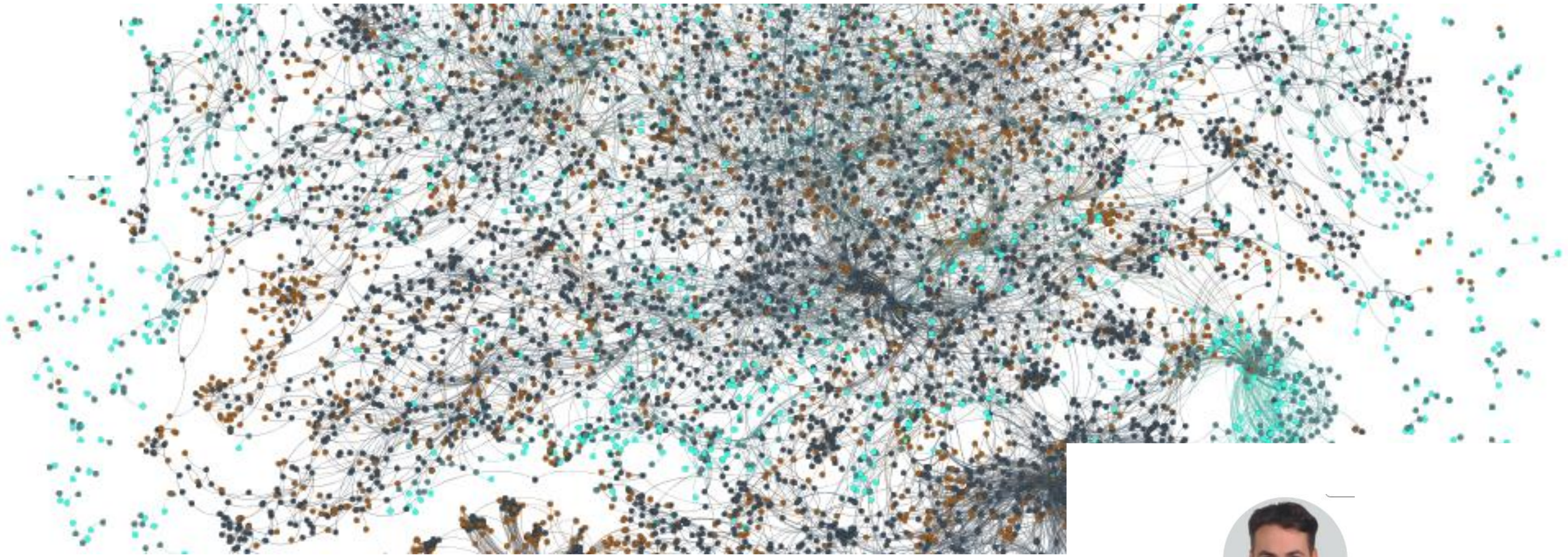
ETL, Data Management and Data Lineage

a.k.a. when life gives metadata, make graph analysis out of it



Luca Perrozzi

10.11.2020

DISCLAIMER: the humor contained in this talk is based on real life experience but is not meant to provide a general representation of the professional figures involved! ☺

The quest for data warehouses and the problems STILL to be solved

- Large financial institutions, as every other organization nowadays, produce and own **large amount of data that need to be used to take better and data-driven business decisions**

- **The technical problems implied are typically complex and challenging to solve**:
  - Fast growing data O(PB)
  - Legacy and on-premise architectures (COBOL is still a thing!)
  - Large variety of etherogeneous systems to be integrated
  - Strict security levels must be enforced

The quest for data warehouses and the problems STILL to be solved

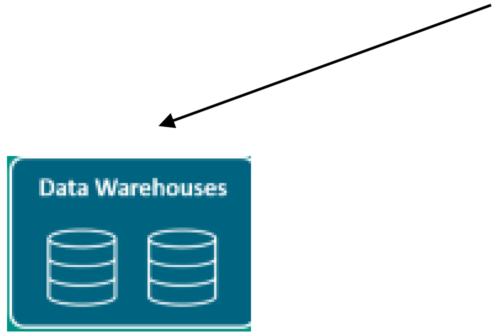
- In the last years we witnessed an **explosion of tools and architectures to address these technical problems**, and solutions are becoming more and more standardized, robust and scalable
  - Just think of Hadoop, plethora of OLTP/OLAP DBs, Cloud vendors, …

- The **real problems** (always) arise when the **technical solutions and operations need to co-exist with the business** expectations, evolving (conflicting) requirements and top-down decisions
  - Ever tried to convert business language to technical solution and viceversa?
  - Technical and business priorities are hardly overlapping
  - Communication is always an issue, especially in large organizations
  - Difficult trade-offs of **speed vs security vs costs vs maintainability**
  - «One tool to solve all problems» un-realistic promises

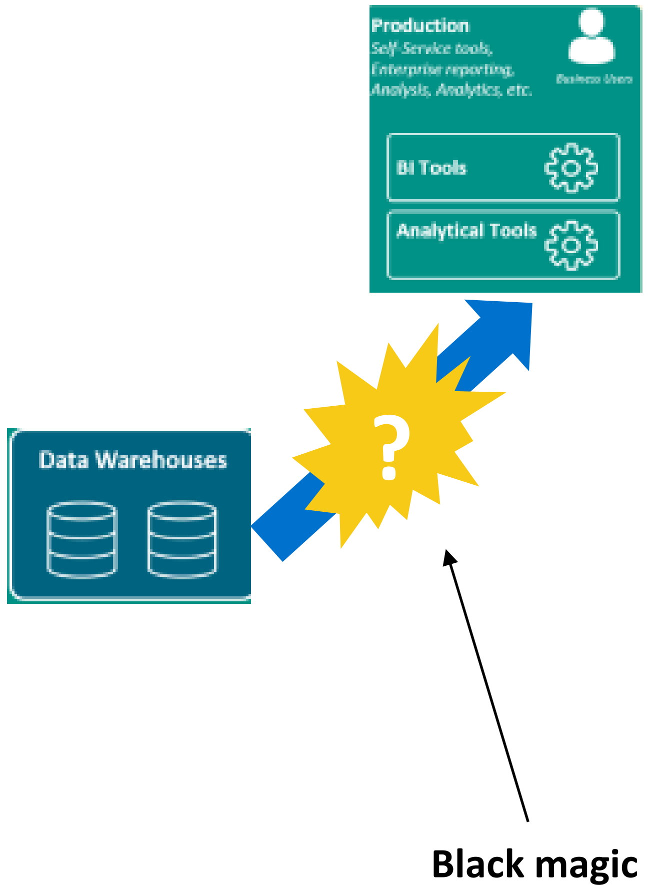
This is what "business" typically knows

We have a cool branded data warehouse solution!
We can conquer the world!

- Integrates several functionalities
- Promises to solve all the problems in one
- Costs a lot (this is why business knows!) ☺

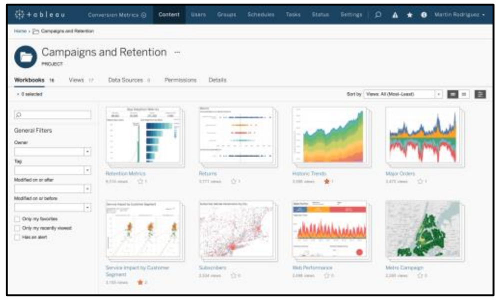


This is what “business” sees

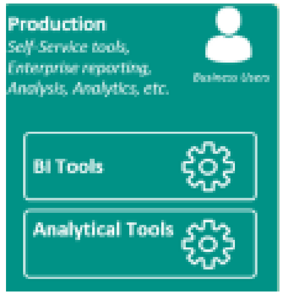


This is what “business” asks

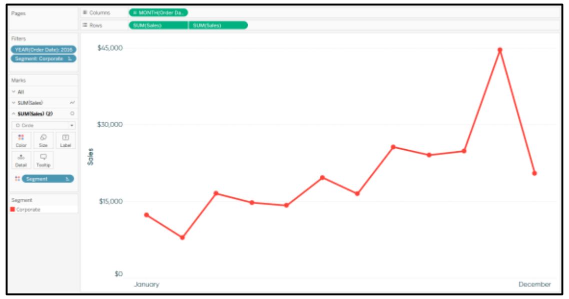
This graph is too complicated!



Qualitative, debatable question


Production
Self-Service tools, Enterprise reporting, Analysis, Analytics, etc.
Business Users
BI Tools
Analytical Tools

This graph is too simple!

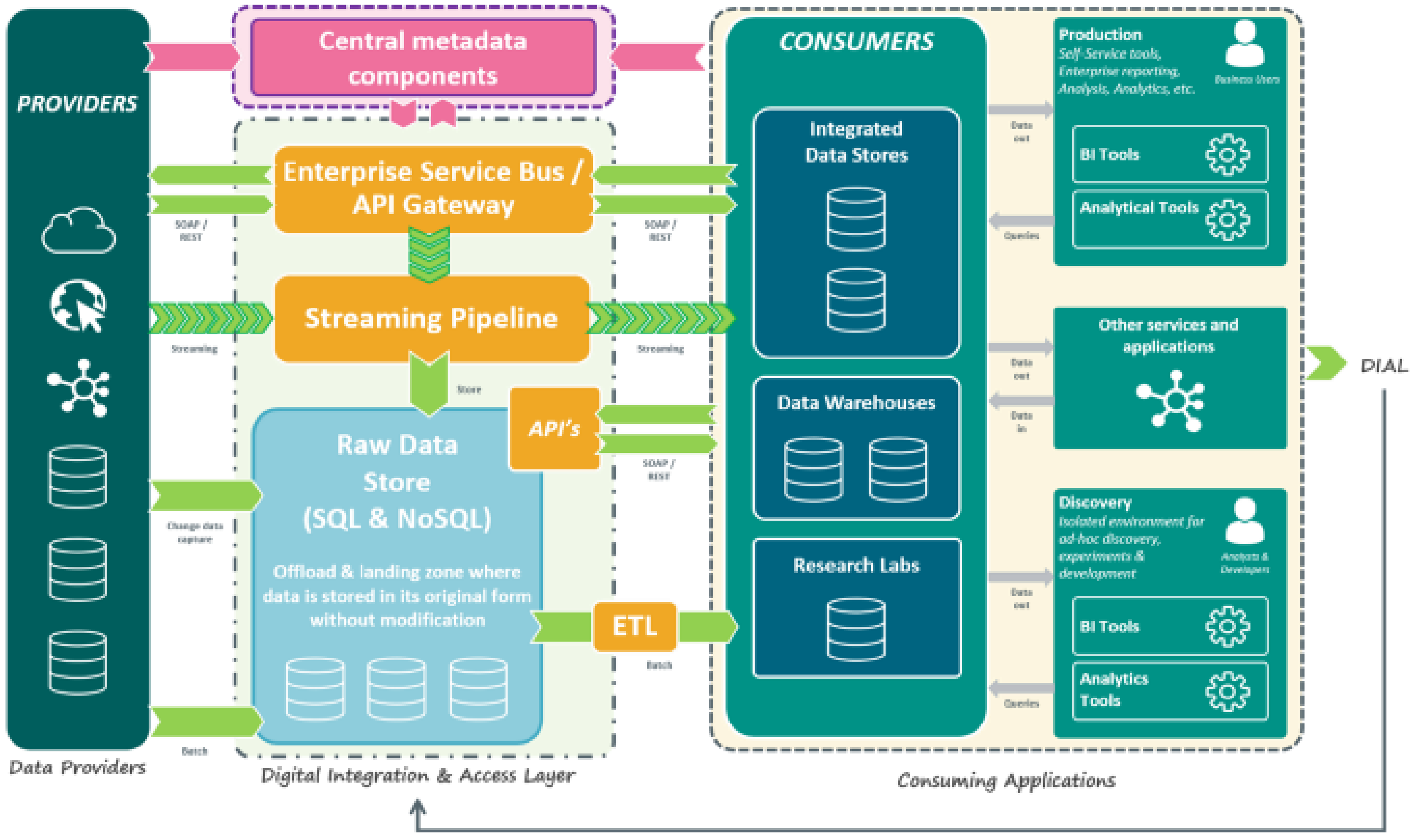


Qualitative, debatable question

Quantitative, important questions

- **Which version of the data did you use?**
- **Did you fix the upstream bug?**
- **Is the data updated?**
- **Is the data legally approved?**
- **Can we leverage more data?**
- **Can we re-use data from other projects?**
- **Who else is using this data?**

Good luck in finding out… (still a simplified version!)
How to make sense of such a complicated chain?



Central metadata components
PROVIDERS
Enterprise Service Bus / API Gateway
SOAP / REST
Streaming Pipeline
Streaming
Store
API's
Raw Data Store (SQL & NoSQL)
Offload & landing zone where data is stored in its original form without modification
Change data capture
Batch
ETL
CONSUMERS
Integrated Data Stores
Data Warehouses
Research Labs
Data out
Queries
Data in
Production
Self-Service tools, Enterprise reporting, Analysis, Analytics, etc.
Business Users
BI Tools
Analytical Tools
Other services and applications
Discovery
Isolated environment for ad-hoc discovery, experiments & development
Analysts & Developers
Analytics Tools
DIAL
Data Providers
Digital Integration & Access Layer
Consuming Applications

Image source: https://medium.com/abn-amro-developer/abn-amros-data-integration-architecture-f33506a211c3

Answer: use Metadata! (indicative, minimal list)

Technical

- **Data source properties**
    - Platform (CH, R.O.W.)
    - Sourcing team
    - Version
- **Environments (PROD, PTA)**
    - Type of feed (interface)
    - Landing folder
    - Staging folder
- **Data Warehouse attributes**
    - Folder
- **Batch processor**
    - Cron schedule
    - Batch group
    - Batch timeout
- **Landing file format**
- **Staging file format**
- **Data source checks**
    - Check definition
- **Datasets properties**
    - Name
    - File pattern
    - Obfuscation details
    - Ingestion type
    - Versions
    - Dataset checks
        - Check definitions
    - Attributes
        - Name
        - Type

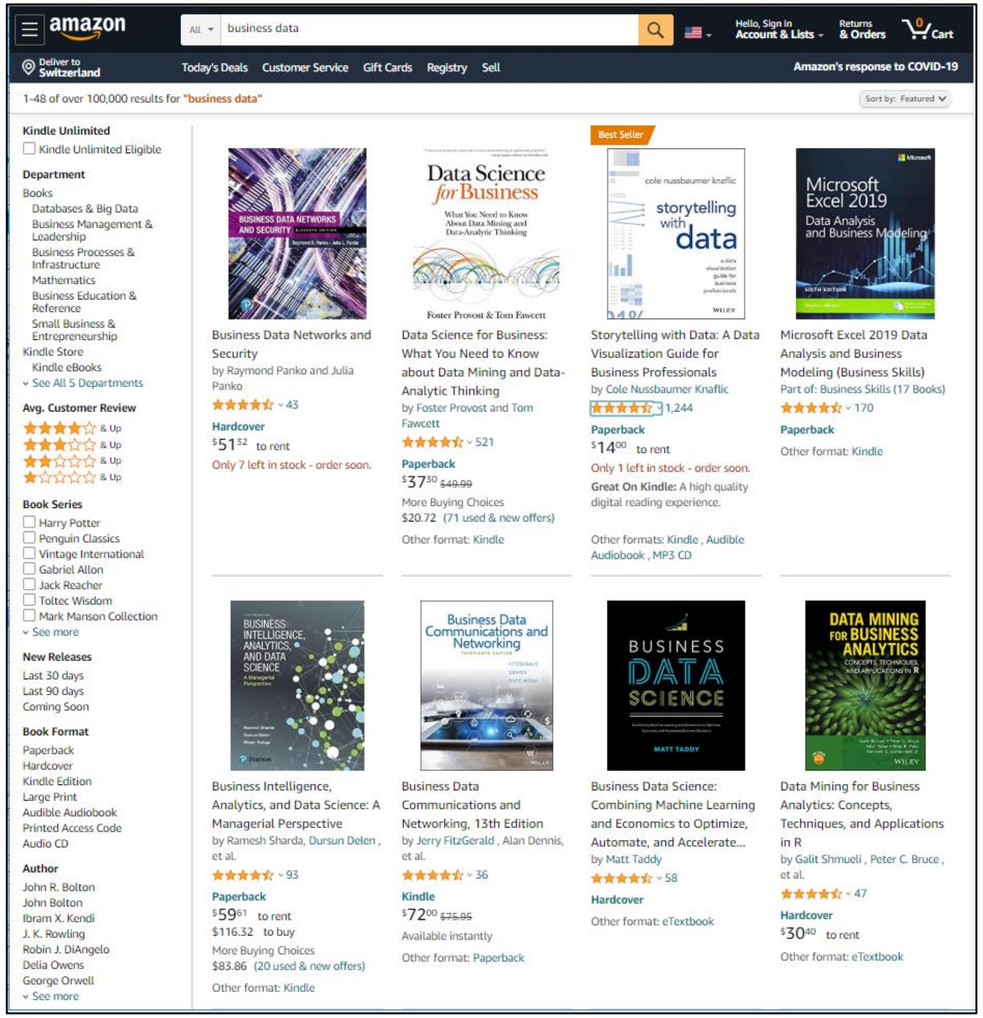
Business

- **Data source**
    - Owner ID
    - Upstream application
    - Description
- **Initial requester**
    - Project
    - Business owner
- **Data category (GDPR)**
- **Service Level Agreement (SLA)**
    - Periodicity
    - Processing time
    - Reaction time
- **Data scope**
    - Division
    - Location
- **Originating/Sending/Receiving legal entities**
- **Delivery type**
- **Approvals**
    - Approval type
    - Approver
    - Date
- **Contacts**
    - Contact type
    - Contact name
    - Email
- **Data protection**
- **Datasets**
    - CID Class
    - Attributes

Operational (monitoring purpose)

- **Batch name**
- **Execution id**
- **Execution start**
- **Execution end**
- **Processing Stages**
    - Stage start
    - Stage end
    - Stage result
    - Stage error message
- **Statistics**
    - Files statistics
    - Data statistics
- **File controls**
    - Control type
    - Control timestamp
    - Control result
    - Error message
    - Execution stage
- **Datasets**
    - Dataset name
    - Controls
        - Control type
        - Control timestamp
        - Control result
        - Error message

Dreaming about (meta)data representation…



amazon | All | business data | Hello, Sign in Account & Lists | Returns & Orders | 0 Cart

Deliver to Switzerland | Today's Deals | Customer Service | Gift Cards | Registry | Sell | Amazon's response to COVID-19

1-48 of over 100,000 results for "business data" | Sort by: Featured

Kindle Unlimited
- Kindle Unlimited Eligible

Department
Books
- Databases & Big Data
- Business Management & Leadership
- Business Processes & Infrastructure
- Mathematics
- Business Education & Reference
- Small Business & Entrepreneurship

Kindle Store
- Kindle eBooks
- See All 5 Departments

Avg. Customer Review
- ★★★★☆ & Up
- ★★★☆☆ & Up
- ★★☆☆☆ & Up
- ★☆☆☆☆ & Up

Book Series
- Harry Potter
- Penguin Classics
- Vintage International
- Gabriel Allon
- Jack Reacher
- Toltec Wisdom
- Mark Manson Collection
- See more

New Releases
- Last 30 days
- Last 90 days
- Coming Soon

Book Format
- Paperback
- Hardcover
- Kindle Edition
- Large Print
- Audible Audiobook
- Printed Access Code
- Audio CD

Author
- John R. Bolton
- John Bolton
- Ibram X. Kendi
- J. K. Rowling
- Robin J. DiAngelo
- Delia Owens
- George Orwell
- See more

Business Data Networks and Security
by Raymond Panko and Julia Panko
★★★★☆ 43
Hardcover
$51.32 to rent
Only 7 left in stock - order soon.

Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking
by Foster Provost and Tom Fawcett
★★★★☆ 521
Paperback
$37.50 ~~$49.99~~
More Buying Choices
$20.72 (71 used & new offers)
Other format: Kindle

Best Seller
Storytelling with Data: A Data Visualization Guide for Business Professionals
by Cole Nussbaumer Knaflic
★★★★☆ 1,244
Paperback
$14.00 to rent
Only 1 left in stock - order soon.
Great On Kindle: A high quality digital reading experience.
Other formats: Kindle , Audible Audiobook , MP3 CD

Microsoft Excel 2019 Data Analysis and Business Modeling (Business Skills)
Part of: Business Skills (17 Books)
★★★★☆ 170
Paperback
Other format: Kindle

Business Intelligence, Analytics, and Data Science: A Managerial Perspective
by Ramesh Sharda, Dursun Delen , et al.
★★★★☆ 93
Paperback
$59.61 to rent
$116.32 to buy
More Buying Choices
$83.86 (20 used & new offers)
Other format: Kindle

Business Data Communications and Networking, 13th Edition
by Jerry FitzGerald , Alan Dennis, et al.
★★★★☆ 36
Kindle
$72.00 ~~$75.95~~
Available instantly
Other format: Paperback

Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate...
by Matt Taddy
★★★★☆ 58
Hardcover
Other format: eTextbook

Data Mining for Business Analytics: Concepts, Techniques, and Applications in R
by Galit Shmueli , Peter C. Bruce , et al.
★★★★☆ 47
Hardcover
$30.40 to rent
Other format: eTextbook

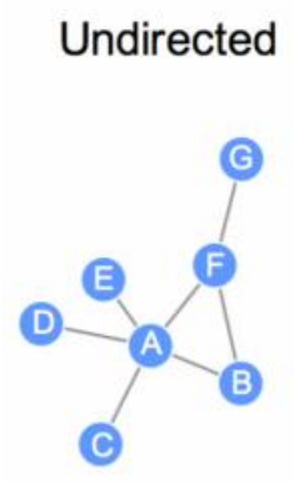
- (meta)Data is perfectly **categorized**
- Technical, Business and Operational (meta)data **harmonized** and stored in a single place
- (meta)Data is **searchable**
- (meta)Data is available in **different formats**
- **Relations/similarities** between (meta)data are captured (more in the next slides)
- **Detailed descriptions** available to understand data
- **Quality** of the data is captured

How can we reach a similar level of representation?
A good data management model and corresponding metadata collection is the key… and many (long forgotten) tools are available on the market!

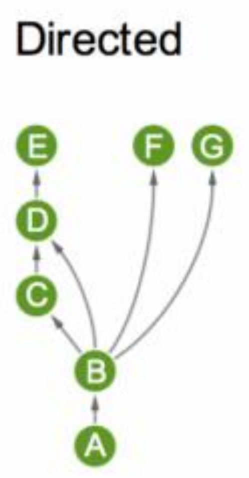
All this might seem quite boring, so let’s focus on a fun aspect: Data Lineage

- Data Lineage is a way of representing relations between datasets making use of graph analysis

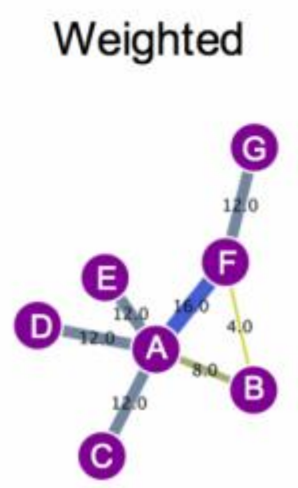
- Graphs come with nodes (letters in the sketches below) and edges (lines in the sketches)
- Graphs can have different edge flavors (undirected, directed, weighted)



Social media platforms
(Facebook, Twitter, Linkedin, …)



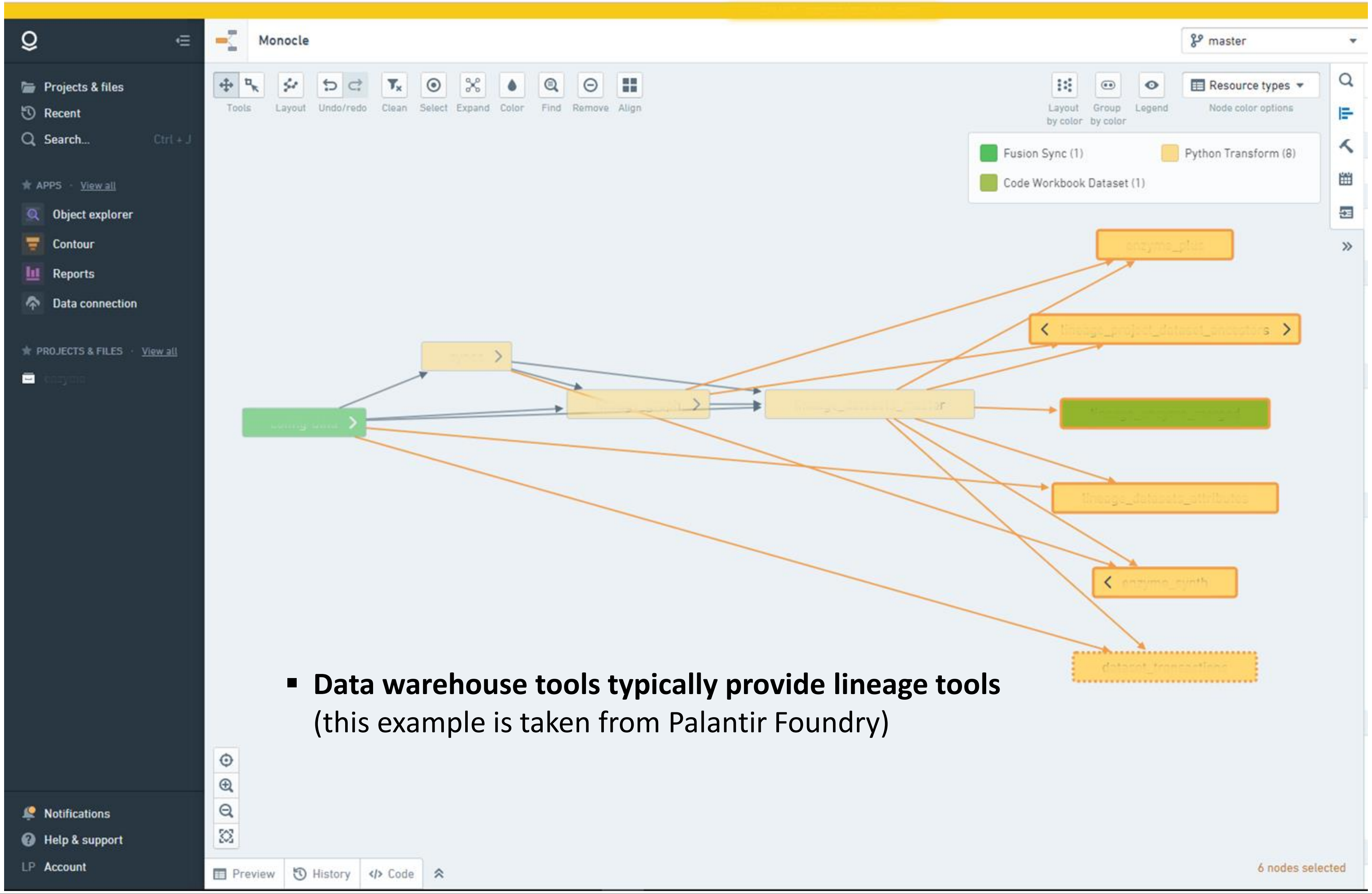
Data platforms,
Batch processing systems,
Cooking (!)



GPS navigation systems
(Google maps, Tom Tom, …)

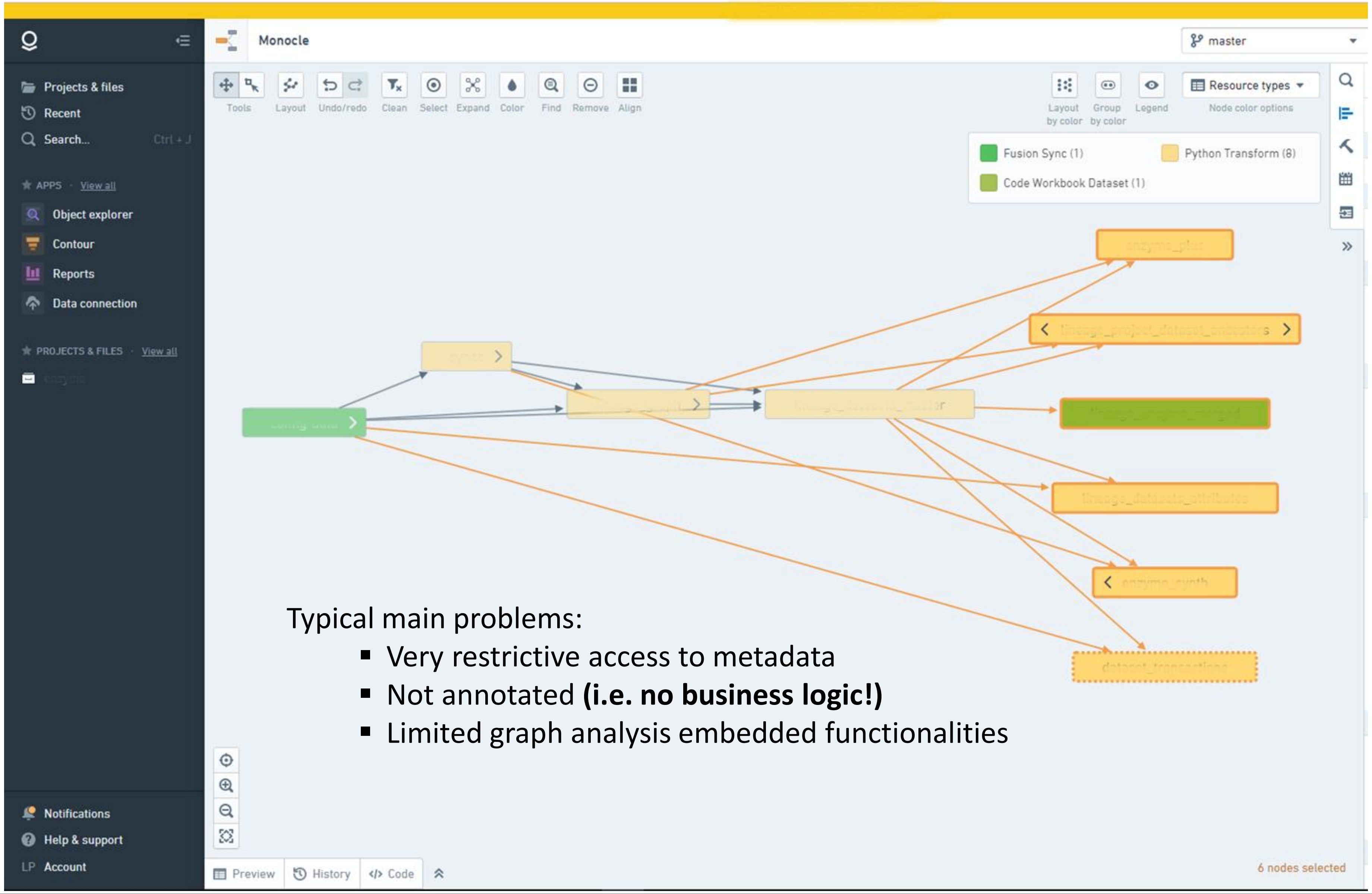
- Data platform graphs are typically «**Directed**» as we have **parent-child relations between nodes**

Typical data warehouse lineage: end of the story?



- **Data warehouse tools typically provide lineage tools** (this example is taken from Palantir Foundry)

Typical data warehouse lineage: what else do we need?

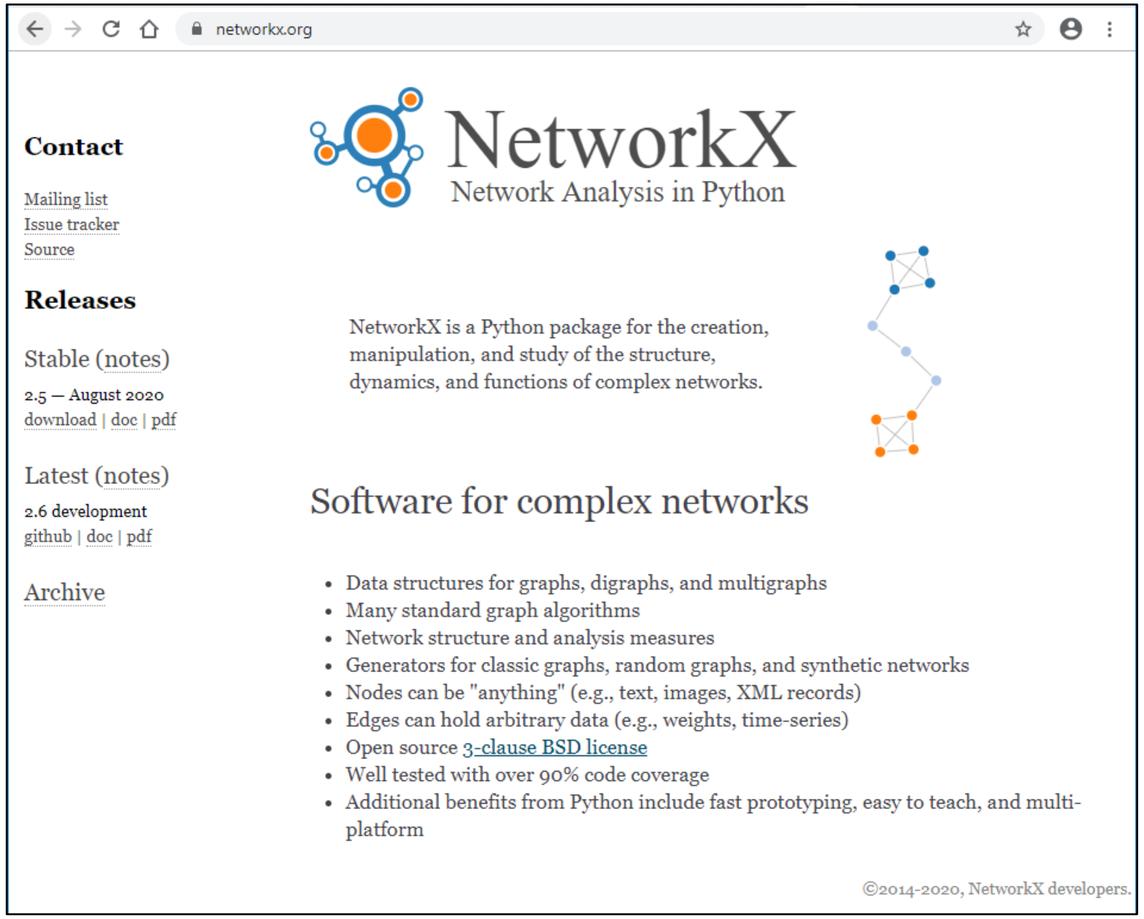


Typical main problems:

- Very restrictive access to metadata
- Not annotated **(i.e. no business logic!)**
- Limited graph analysis embedded functionalities

Annotated lineage: technical solution

- Technical solution: use **networkx** (alternative solution: Java-based Neo4j)
  - An **open source Python package** for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.



networkx.org

Contact

Mailing list
Issue tracker
Source

Releases

Stable (notes)

2.5 — August 2020
download | doc | pdf

Latest (notes)

2.6 development
github | doc | pdf

Archive

NetworkX
Network Analysis in Python

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

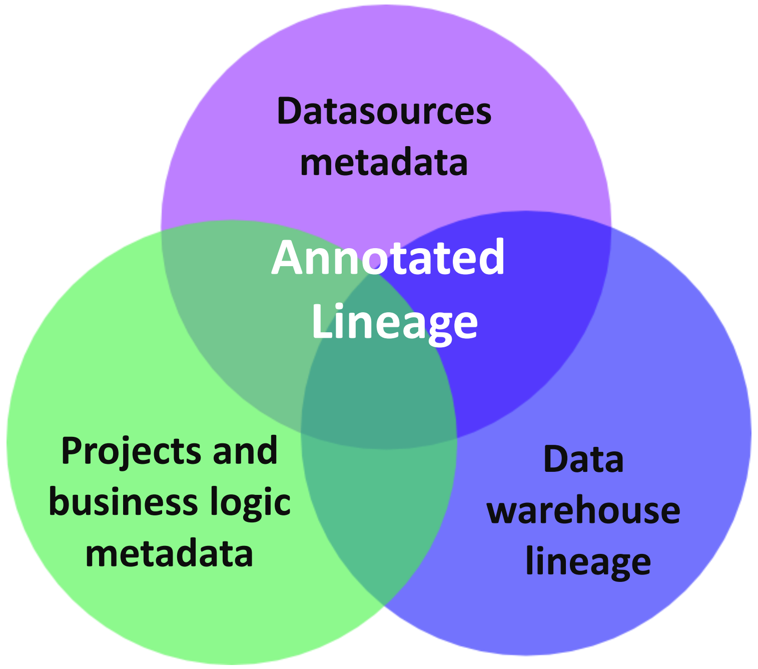
Software for complex networks

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

©2014-2020, NetworkX developers.

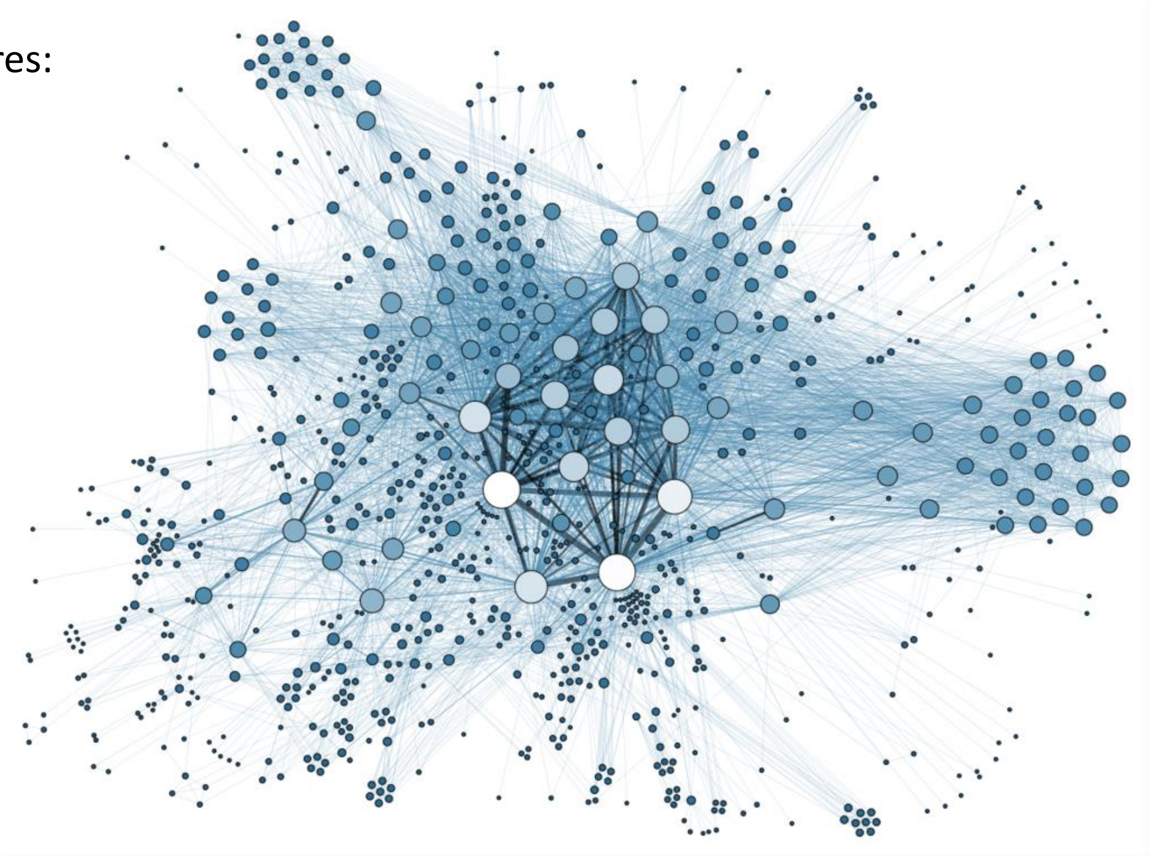
Annotated lineage: know your platform

- **Create and annotate** platform graph from data warehouse API and all the available metadata
  - **Solve ambiguities** and identify documentation errors
  - **Deal with different entities arising from business logic**: datasources, projects, others
  - **Cleanup** and add missing documentation
  - **Report** possible policy violations
  - **Answer business questions** in a fast and effective way



Data warehouse lineage: typical facts and figures

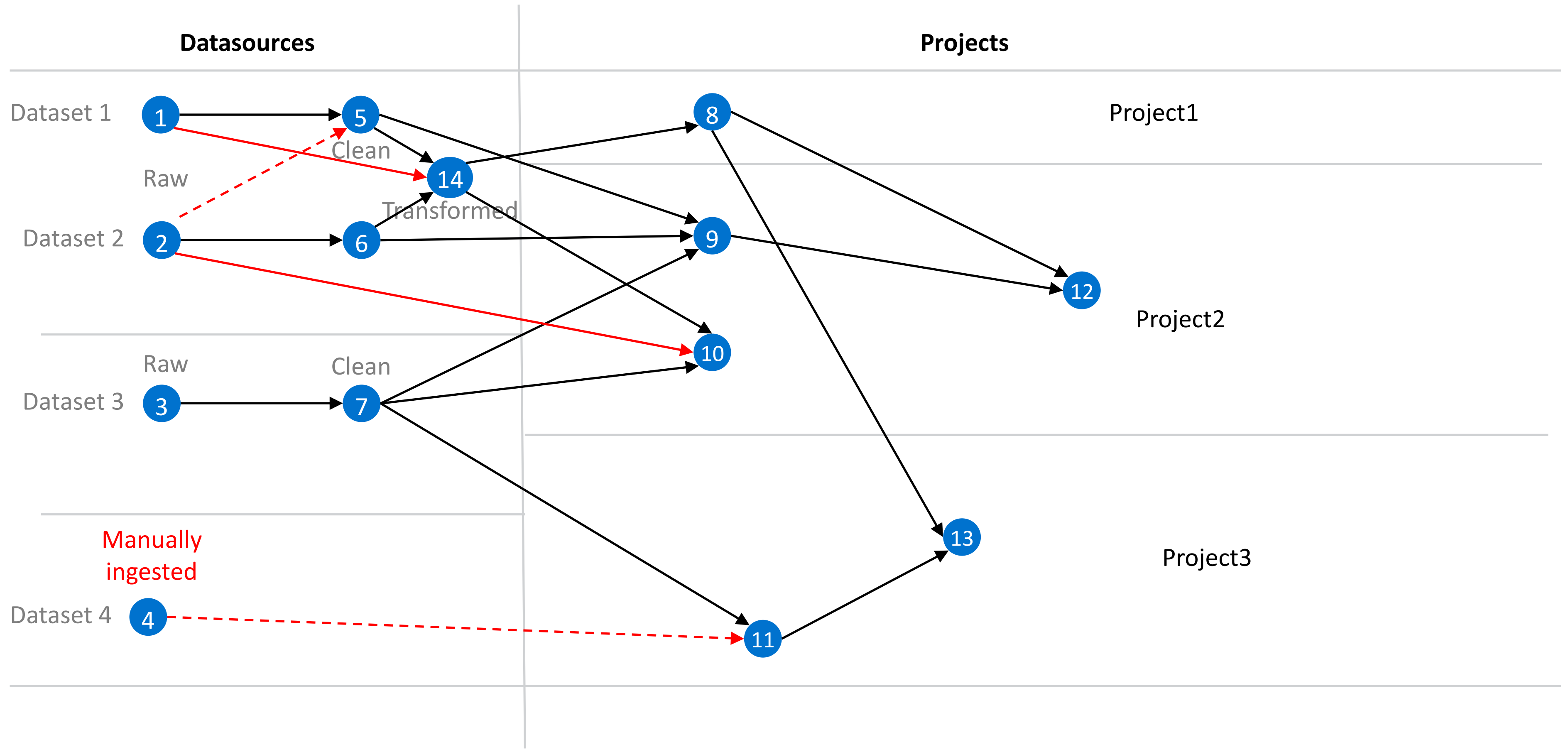
- Typical Data warehouse lineage features:
  - O(100)   datasources
  - O(10)   projects
  - O(10k)   nodes
  - O(100k) direct connections
  - O(10M) indirect connections



NB: each platform (R.O.W., CH) and environment (UAT, PROD) has its own distinct lineage graph!

Image source: https://blog.grakn.ai/life-on-the-edges-september-2016-cb4b900dd0a0

Lineage simplified (!) schematics



- Two possible dependency relations for each node:
  - **Forward depencencies**: which other nodes depends on the current node?
  - **Backward depencencies**: which other nodes is the current node depending on?

How to answer Lineage related business questions?

Main ideas implemented to cope with system limitations:

- Create a **«flattened» static representation of the graph using python package pandas** at the expense of duplicating information
  - One entry for each pair of connected nodes, either directly or indirectly
  - Updated daily

O(100k) rows

Parent dataset	Child dataset	Direct connection
A	B	TRUE
B	C	TRUE
A	D	TRUE

Parent dataset	Child dataset	Direct connection
A	B	TRUE
A	**C**	**FALSE**
B	C	TRUE
A	D	TRUE

O(10M) rows

- Indentify most common cases and **build reports** in excel or similar format for easy consumption from stakeholders
  - Provide overview and detailed versions

- Collect **lineage** for each platform (R.O.W., CH) and environment (UAT, PROD) **in one place**
  - Use **sharepoint** or similar technology to collect reports, so they can be consulted and searched online in the browser, and can be downloaded on the laptop for more refined analysis

Lineage related business questions and reports

Typical business questions:

1. **Which projects depends on a given datasource?**
    - Report: list dependent projects for each datasource
2. **Business question: which other projects depend on a given project?**
    - Report: for each project, list other projects dependending on it
3. **Which datasources and other projects is a given project depending on?**
    - Report: list datasources and other projects dependencies for each project

Policy violations reports:

1. **Ingested but unused datasets**
2. **Stale datasources without ingestions in the last 3 months**
3. Raw datasets directly consumed by projects
4. Direct ingestion of datasources into project folders
5. Manually ingested datasets
6. Datasource depending on other datasources
7. Datasources using non-standard naming conventions

…

Lineage and networkx DEMO

The material used in this demo can be found here:
https://github.com/perrozzi/networkx_example

Conclusions

- Nothing new under the sun, but worth repeating ☺
- Data warehouses solve a whole lot of technical problems
- As usual, the **real problems** arise when the **technical solutions and operations need to co-exist with the business** expectations, evolving (conflicting) requirements and top-down decisions
- Data management and metadata are typically the most mistreated and disregarded information in this context
  - Still without them the technical solution will solve some problems and create (many) others
- Data management requires the coordination of several interested parties, both from the technical and business side
- Graph analysis and lineage allow an in-depth knowledge of data platform and can be extremely useful to answer business questions
  - And graph analysis is fun to implement and play with!