Storing & Extracting Data

Learning objectives

- Exposure to the most common data storage ecosystems.
- Learn the most common challenges around ingestion, storage, and extraction of data.
- Practice loading different types of datasources in Tableau.

Data Trends

Back in 70's, when the first databases were invented, storing data was absurdly expensive.

That has changed and consequently we are capturing exponential amounts of data.

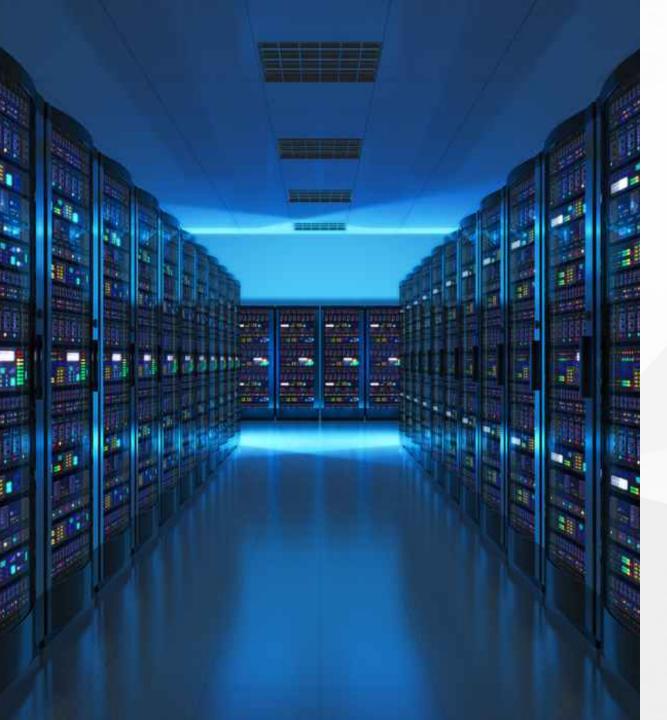
30 0.1 0.09 25 0.08 0.07 20 0.06 0.05 15 0.04 10 0.03 0.02 0.01 2011 2012 2013 2014 2015 2016 2017 2010 Global data created in zettabytes Cost of 1 terabyte of storage, in \$ (rhs)

Figure 3: Costs of storage and global data availability, 2009-2017

Source: Reinsel, Gantz and Rydning (2017); Klein (2017). One zettabyte is equal to one billion terabytes.

Spreadsheet Galore

Imagine your organization wants to store operational data in a spreadsheet (e.g. Excel). Do you think this is a good idea or not?



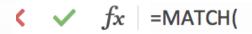
Where is the data?

Data is generally scattered:

- Flat files
- Databases
- Data Warehouses
- Data Lakes
- Source systems
- APIs

Should you store your data in a spreadsheet?

Imagine you work for an organization that wants to start capturing their data. What would be the pros and the cons of storing their data in a spreadsheet (e.g. Excel)?



В	С	D	Е	F	G	
					9	
Name	Last name	Gender				
Sophia	Clark	f				
Samantha	Harris	f				
Emma	Morgan	f				
Isabella	Rodriguez	f				
Olivia	Turner	f				
Elizabeth	Watson	f	=MATCH(
Jacob	Barnes	m	MATCH(Ic	okup_value;	lookup_arra	y;
Robert	Brooks	m				Ė
Oliver	Green	m				
Harry	Howard	m				
Kyle	James	m				
Oscar	Johnson	m				
James	Martin	m				
James	Parker	m				
Rhys	Philips	m				
Joseph	Ramirez	m				
Liam	Reed	m				
Damien	Robinson	m				
Michael	Ross	m				
Jack	Scott	m				
Charlie	Thomas	m				
Daniel	Williams	m			5	
					5	

Flat Files

Common formats include .csv,

- .txt, .xlsx, .json, xml,
- .avro , .parquet .
- + Flexibility.
- Collaboration.
- Security.
- Scalability.





















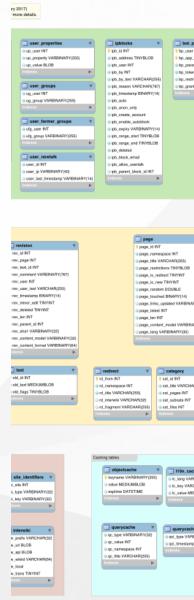




Databases

Databases are an attempt to professionalize data storage. They are composed of spreadsheet-like tables that are related to each other.

- + Collaboration.
- + Security.
- Flexiblity.
- Scalability.
- Tabular.





pl_title VARCHAR(258

ivit from INT

el_from INT

o el to BLOB

o el_index BLOB

II_lang VARBINARY(20)

◆ II_skie VARCHAR(256)

int polic VARRIN

Ivil. 18th VARCHAR(255)





al_user INT

o oi user text VARCHARIZES

oi_timestamp BINARY(14)

 oi_metadata MEDIUMBLOB O oi media type ENUM(...)

a di major mime ENUM...

a oi, minor, mime VARBINA

ai_deleted TINYINT

oi_shu1 VARBINARY

rc. timestamp VAR

rc. namespace INT

o rouminor TIMYIMT

rc_bat TINYINT

rs. new TINYINI

rc_this_oldid INT

rc_last_oldid INT

rc_type TINYINT

ro_source VARCHAR(

rc_patrolled TINYINI

гс_ір УАПВІМАЛУ(4

ro_old_lon INT

rc_new_ien INT rc_deleted TINYIN rc_logid INT rc_log_type YARBINARY(256)

ro. our..id INT

rc #8e VARCHARI255)

on: user text VARCHAR(25)

ro_comment VARCHAR(767

one user INT

vi user N

vi nameso

wi_title VA

O will notificat



ont title VARCH

cat_pages INT

o cat_subcats INT

and files INT

pp_app_id TEXT

bp_taken TEXT

pp_password TEXT

bp_restrictions TEXT

ne mage INT

GDJB sulav_gg •

page_restri

o pr. page INT

gr_user INT

or type VARBINARYIN

or level VARBINARY (6)

pr_casoade TINYINT

pr_expiry VARBINA

__ protected_titles

pt user INT

o pt_title VARCHAR(256)

pt reason VARCHAR(767

pt. timestamp BINARY(14)

o pt. expiry VARBINARY(14)

pt_create_perm VARBINATTY)

gr id INT

gg_sortkey FLOA

ap.aropname VAPBINARY(60



i_from_namespace

I to VARCHARIZED

1 from namespace IN

1 namespace B/T

1 Stie VARCHAR(25)

el_to VARCHAR(255)

cl_type ENUM(...

sl_sortkey VARBINARY(230

ol timestamp TIMESTAM!

el_eoliation VARBINARY(3

cl_sortkey_prefix VARCHAR(2)

f_from INT

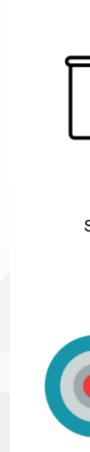
ol from INT

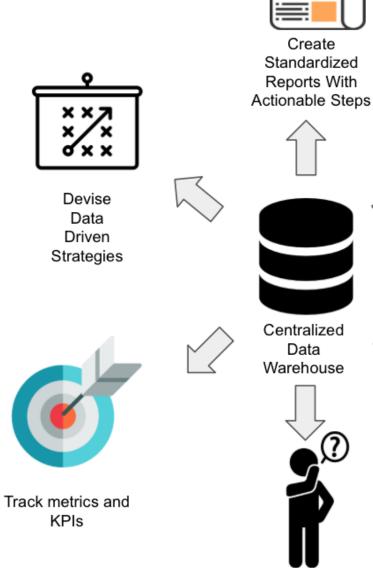


Data Warehouse

Data warehouse look and feel like a database. But they are optimized for analytics (instead of powering an application).

- + Collaboration.
- + Security.
- + Scalability.
- Flexibility.
- Tabular.







Answer Ad-hoc questions

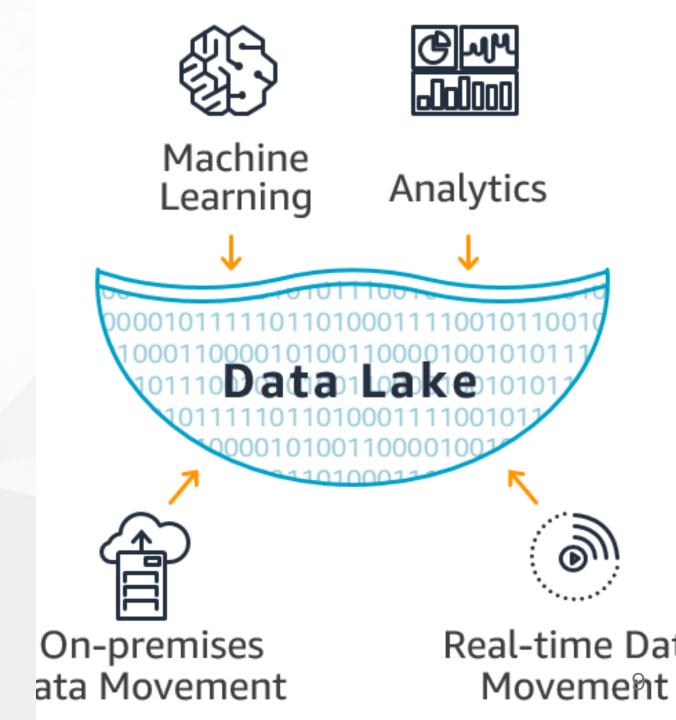
Data

Create

Data Lake

Data lakes are a cheaper, more flexible of data warehouses. Data doesn't need to be tabular or relational anymore. But it can get messy.

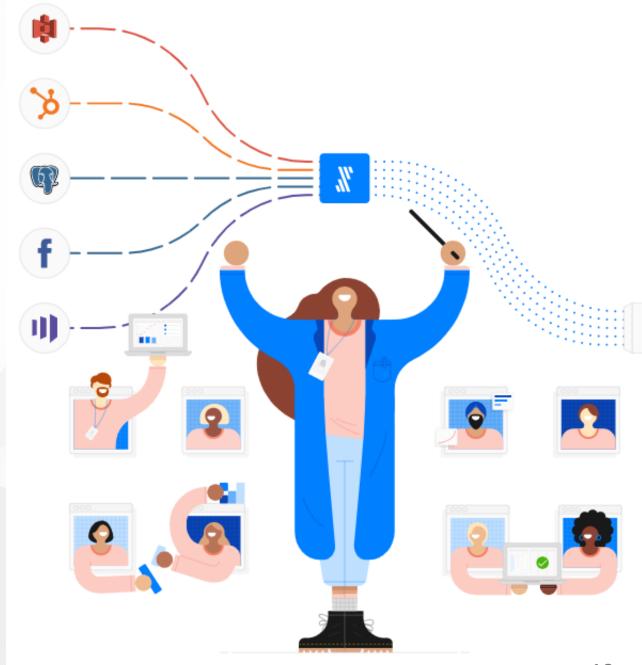
- + Collaboration.
- + Scalability.
- + Flexibility.
- Order.



Source Systems

Source system is any system that captures data. You generally don't want it to live here.

- Flexibility.
- Centralization.



APIs

APIs can be used as a secure interface to allow anybody to query data.

- + Automation.
- + Security.

