DEEP DIVE

Modern Data Pipelines:
Improving Speed, Governance and Analysis

www.dmradio.biz

Featured Speakers

Taylor Brown • 2nd

Co-Founder at Fivetran

San Francisco Bay Area

Eric Kavanagh

eGov Consultant at United Nations

United Nations • Spring Hill College

Austin, Texas Area • 500+

General Electric gets booted from the Dow

by Matt Egan @MattEganCNN

June 19, 2018: 6:14 PM ET

Constraints Drive Design

When conditions change, objectives must

Highway design = commensurate with traffic

No more Moore, massive parallelism better?

Maybe some applications really should die

No time like the present to begin anews!

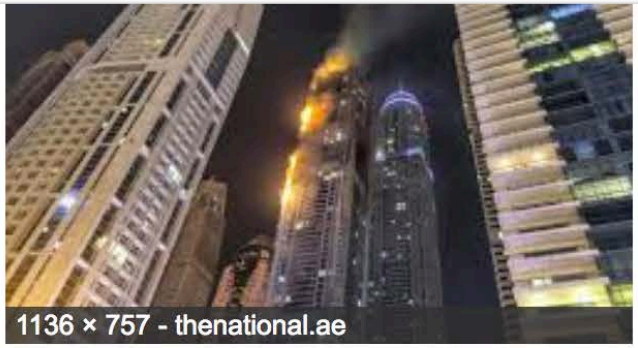
Architecture Matters

Engineering Revolution Enabled Skyscrapers

By Joe Mabel

https://www.google.com/search?q=torch+tower+dubai+fire&tb 120% Search




1136 × 757 - thenational.ae

Dubai Torch tower blaze: residents th…

The National

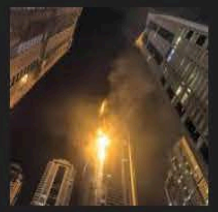
Dubai Torch tower blaze: residents thought it was a false alarm

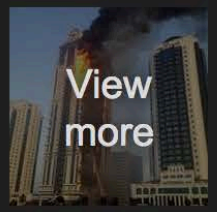
Visit | Save | View saved | Share

Related images:







1136 × 757 - Images may be subject to copyright Learn More

Don’t Forget the Basics: Costs Matter!

- Modern solutions have cost structures too
- Project planning will always be a moving target
- Build in some financial buffers to help ensure long-term success

www.dailymail.co.uk/travel/article-1240729/Burj-Dubai-tallest-building-world-renamed-Burj-Khalifa.html

As the tallest building in the world opened to great fanfair in Dubai yesterday, the struggling emirate was well aware that it owed a big thank you to its oil rich neighbour.

The thanks came in the form of a naming ceremony, Dubai's ruler renamed the previously-known Burj Dubai the Burj Khalifa.

Just last month the tower's namesake and leader of Abu Dhabi, Sheikh Khalifa bin Zayed Al Nahayan, bailed out indebted Dubai to the tune of $10bn - £6.13bn.

© Reuters

High life: The Burj Dubai stands at 160 storeys tall and is the tallest building in the world


Share your plans with
key stakeholders!
Good communication
helps to ensure success!

Process Matters: Continuous Improvement Is Key

- The faster you see value, the more engaged your stakeholders will be
- Create a virtuous circle of improvement by spreading the wealth
- Evangelize success stories; pat your users on the back whenever appropriate
- Starting small is important, but have a long-term plan in mind; this can always change
- “Say yes” whenever possible, even if it’s a tentative “yes” for the near term

In search of the perfect data stack

A brief history of Data warehousing, ETL, BI, and Data Governance

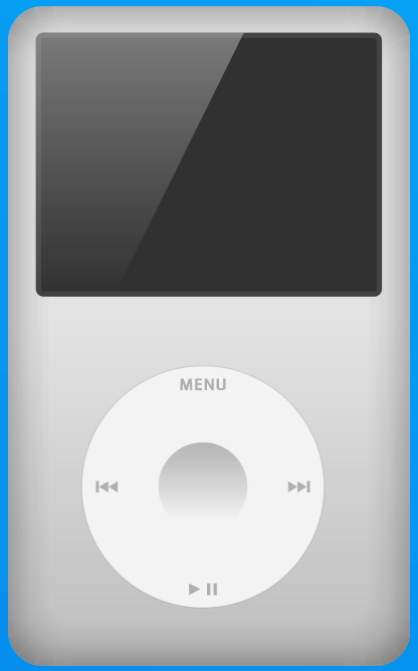
Fivetran

OVERVIEW OF PRESENTATION

- History - looking at trends
- Dates are roughly stated

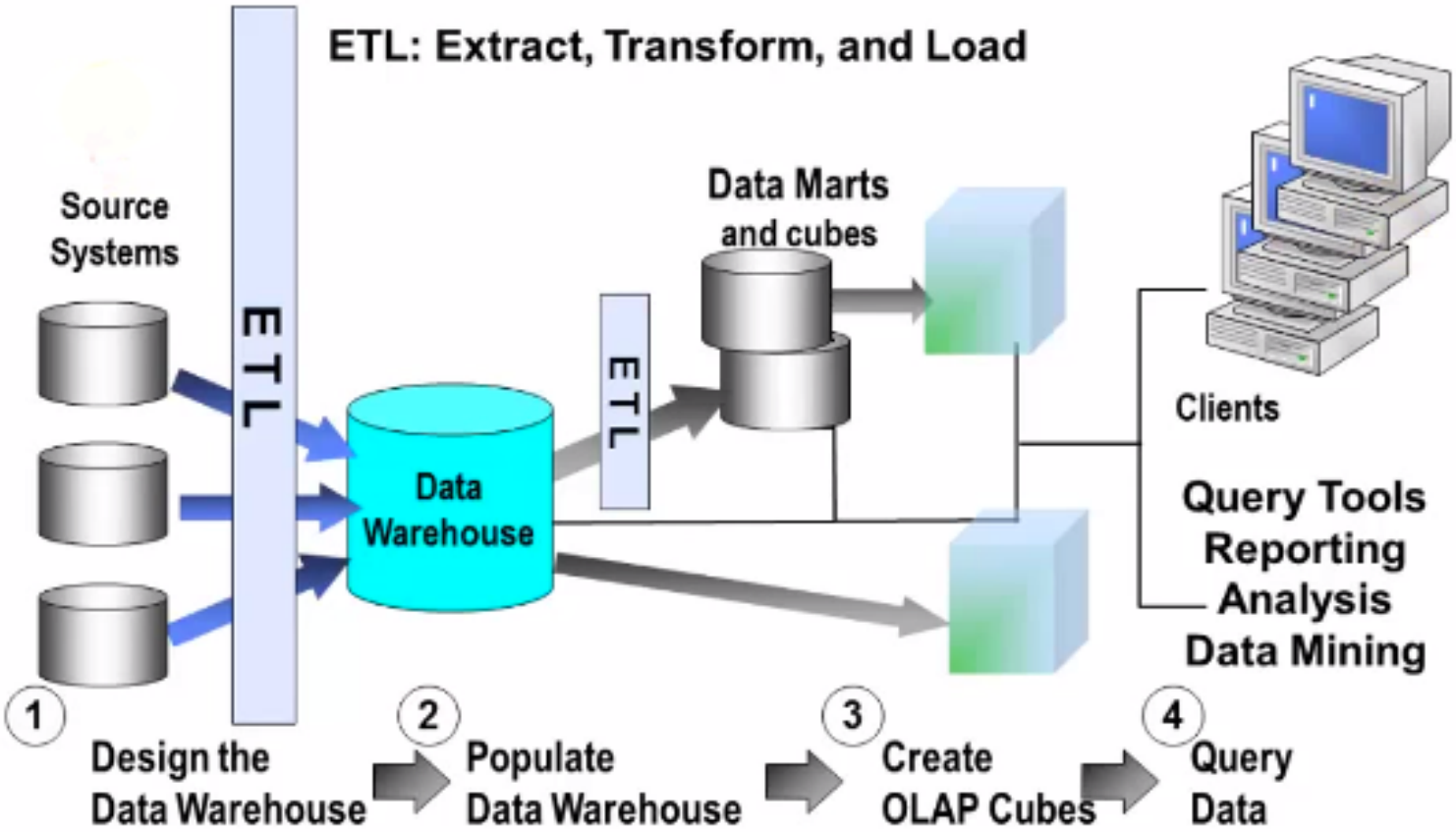
TAYLOR BROWN

COO & Cofounder
taylor@fivetran.com



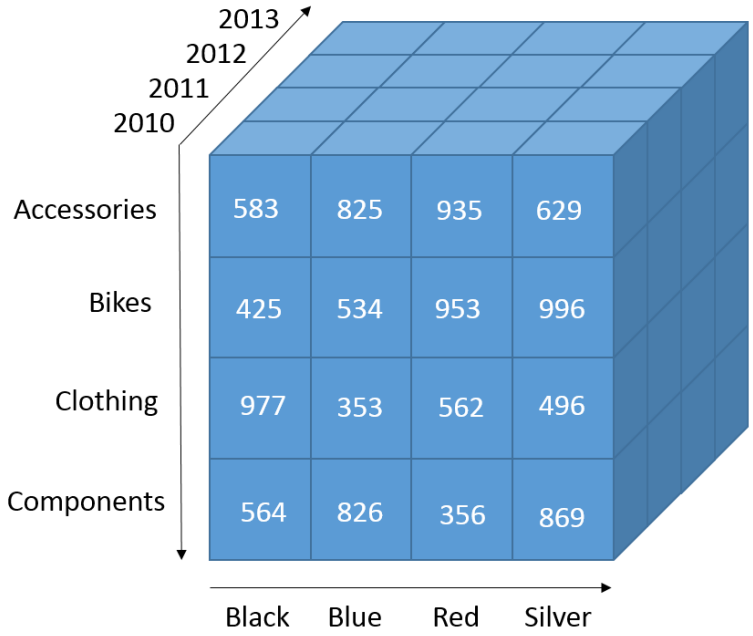
2000’s

2000's Data Stack



http://www.bogotobogo.com/Hadoop/BigData_hadoop_OLTP_vs_OLAP.php

2000's Warehouse - OLAP Cubes



Fast option for analytics vs OLTP Databases

Generally slow and expensive infrastructure

Cost for 1GB = $7.70

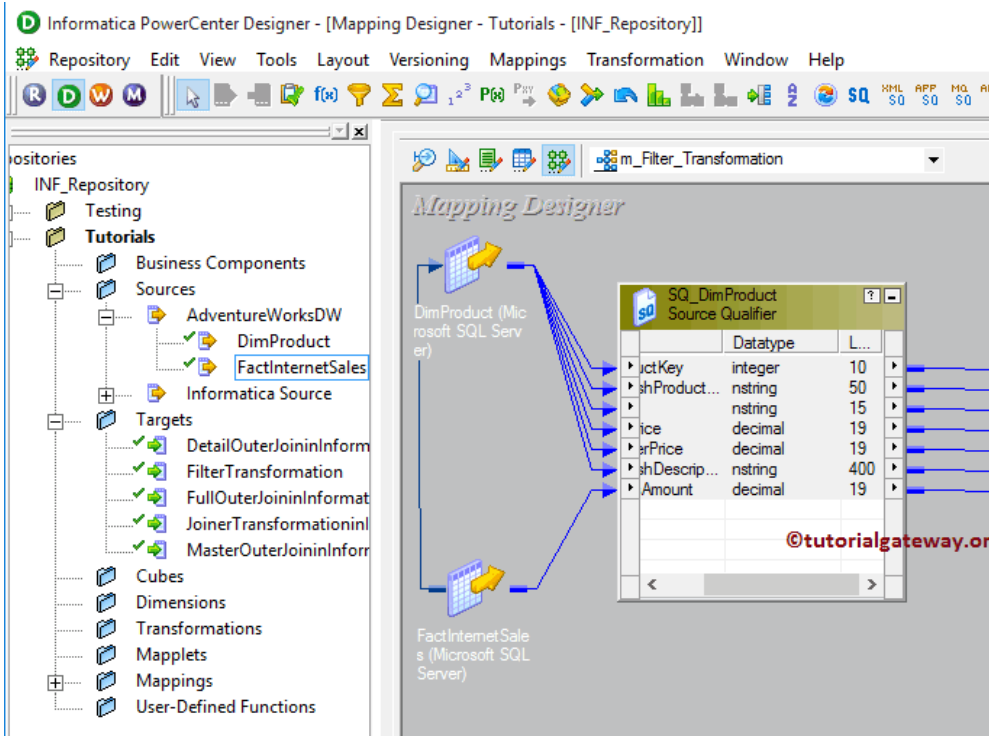
https://community.softwaregrp.com/t5/Big -Data/The -OLAP-cube -is-history/ba -p/231673#.WynTwhJKjMU

2000's Data Pipelines - ETL

Extract, Transformation & Load

Informatica or custom code

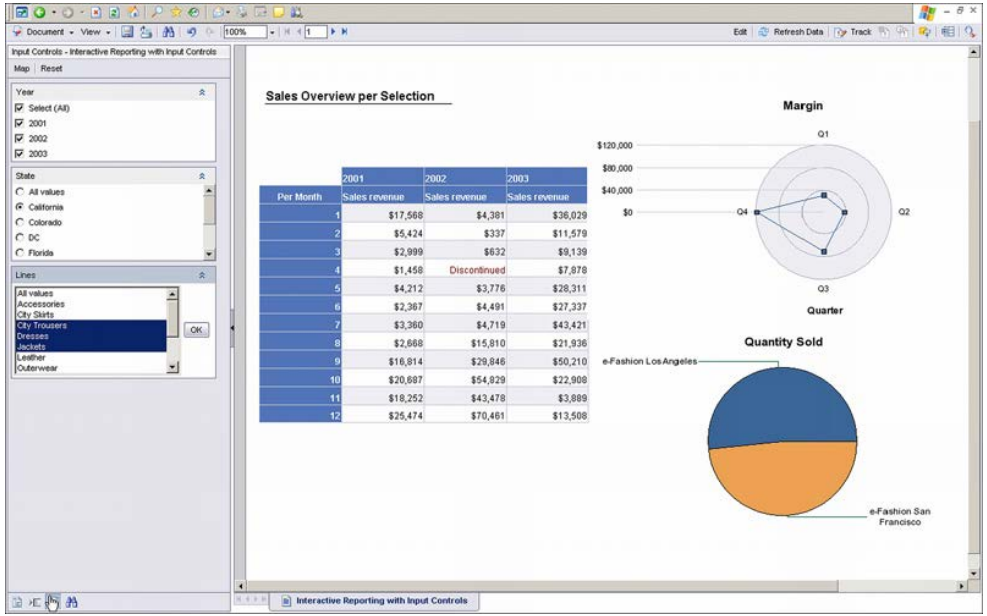
- Heavily customized
- Type, column, table mapping
- Transform data prior to load
- Aggregations performed in pipeline



2000’s Data Governance

- Hardened systems
- Centralized planning
- Good Data Governance

2000's BI Tools



	2001	2002	2003
Per Month	Sales revenue	Sales revenue	Sales revenue
1	$17,568	$4,381	$36,029
2	$5,424	$337	$11,579
3	$2,999	$632	$9,139
4	$1,458	Discontinued	$7,878
5	$4,212	$3,776	$28,311
6	$2,367	$4,491	$27,337
7	$3,360	$4,719	$43,421
8	$2,668	$15,810	$21,936
9	$16,814	$29,846	$50,210
10	$20,687	$54,829	$22,908
11	$18,252	$43,478	$3,889
12	$25,474	$70,461	$13,508

Heavy Monolithic BI tools for Reporting

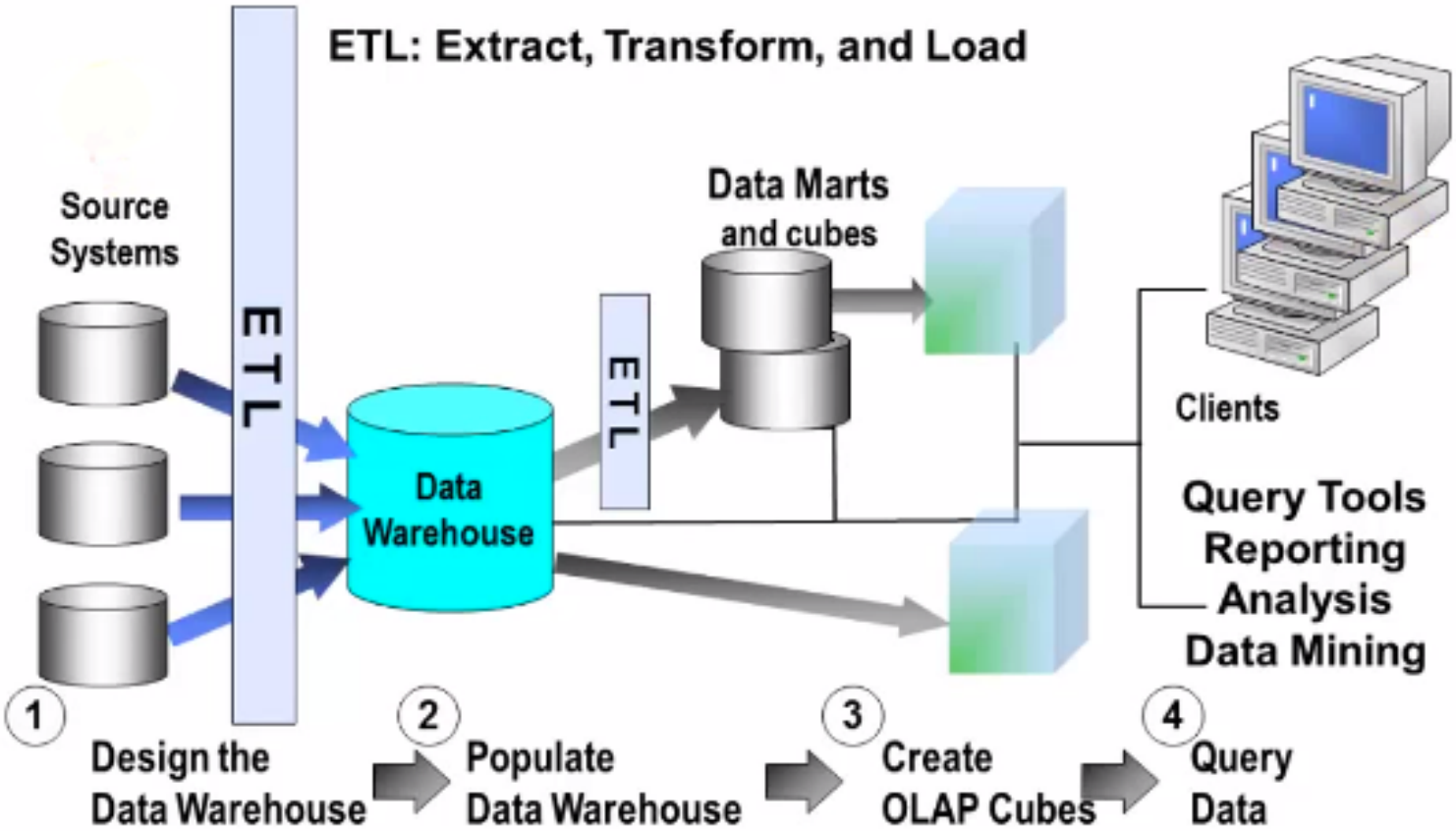
Cognos, Hyperion, Microstrategy

- What happened in past?
- Very Accurate
- Very inflexible.
- Hardened systems
- Months to change

https://www.element61.be/en/resource/sap -business-warehouse-business-objects-front-end-integration-what-available-today

Fivetran

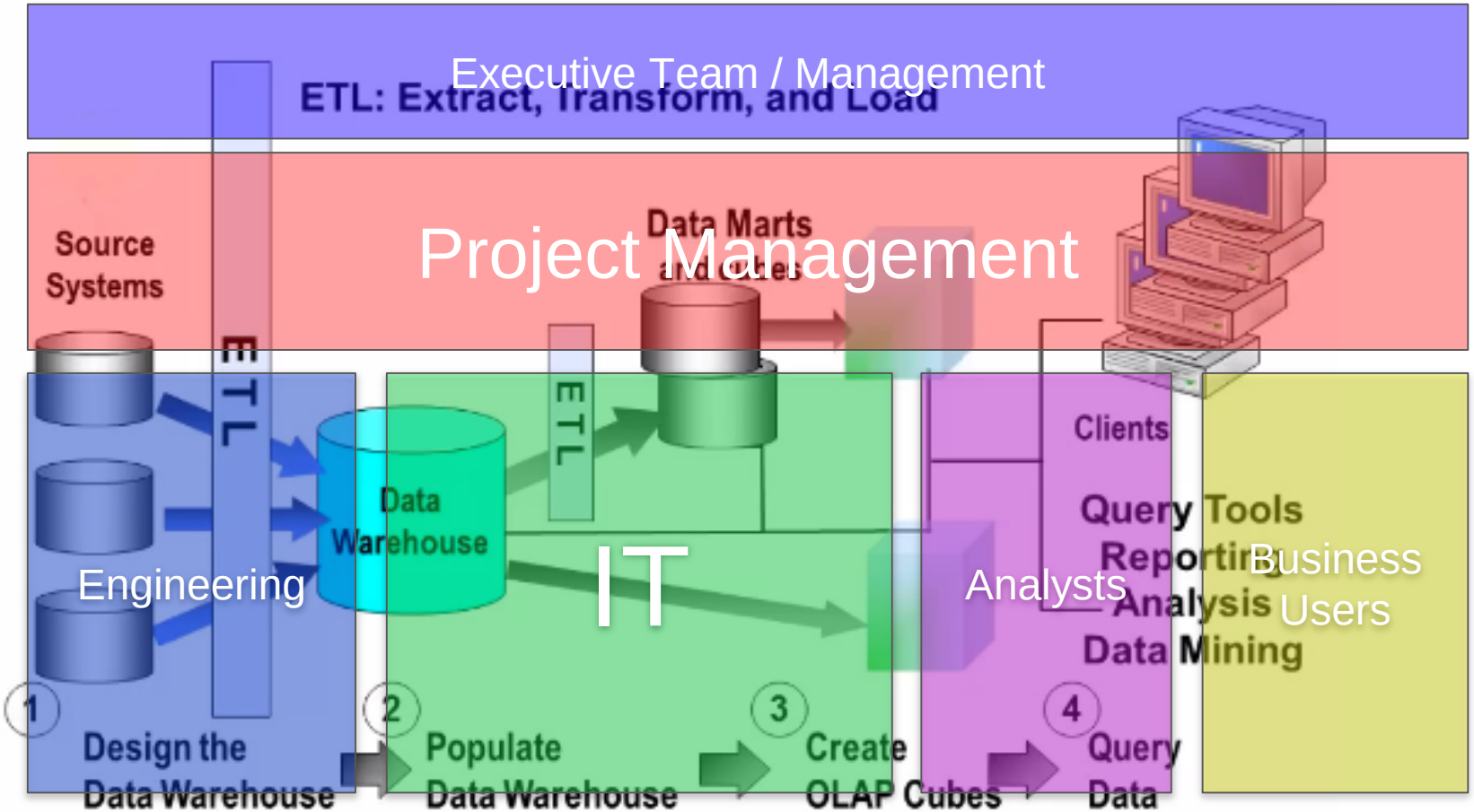
2000s Total Stack = 5+ Tools



http://www.bogotobogo.com/Hadoop/BigData_hadoop_OLTP_vs_OLAP.php

Fivetran

2000s Team Structure = 6+ Teams



http://www.bogotobogo.com/Hadoop/BigData_hadoop_OLTP_vs_OLAP.php


peas
bees
knees
cheese
fleas?

2006

2006’s Challenges with OLAP

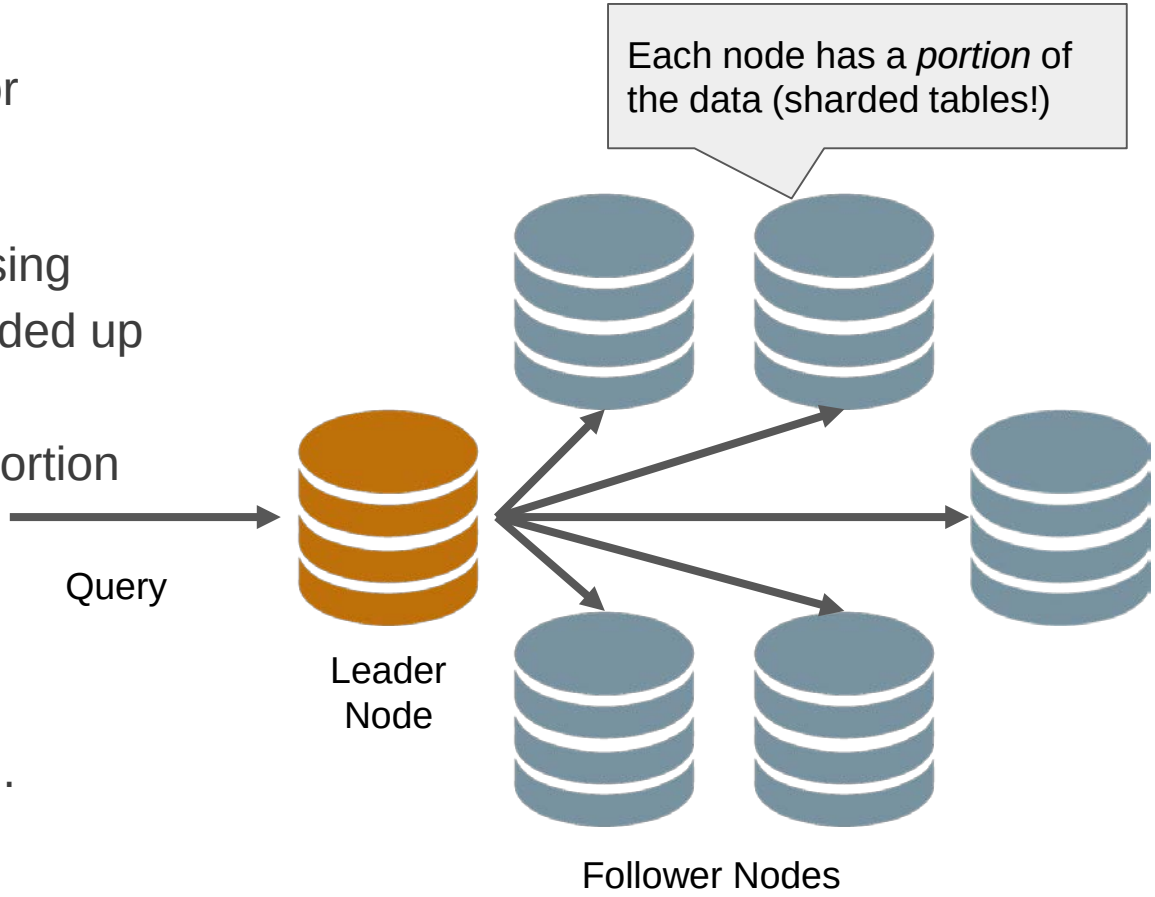
- Data Availability
- Inflexibility
- Speed
- Compromise with end users
- Data volumes

2006's Warehouse - Column Store MMP On -Prem DB

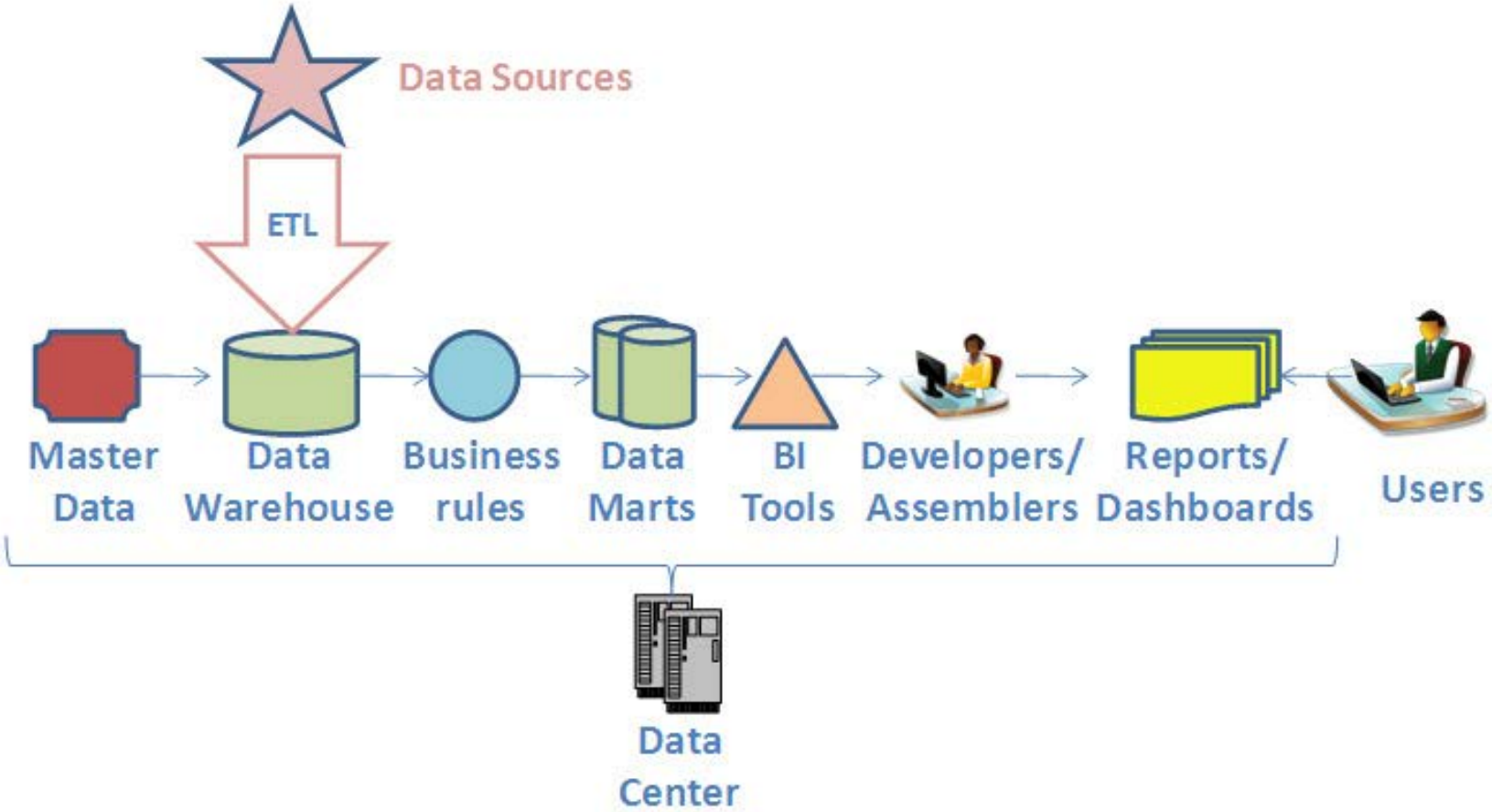
Column-Store designed for analytical queries

Massively Parallel Processing (MPP) - Queries (jobs) divided up between the nodes in the cluster, each one does a portion of the work

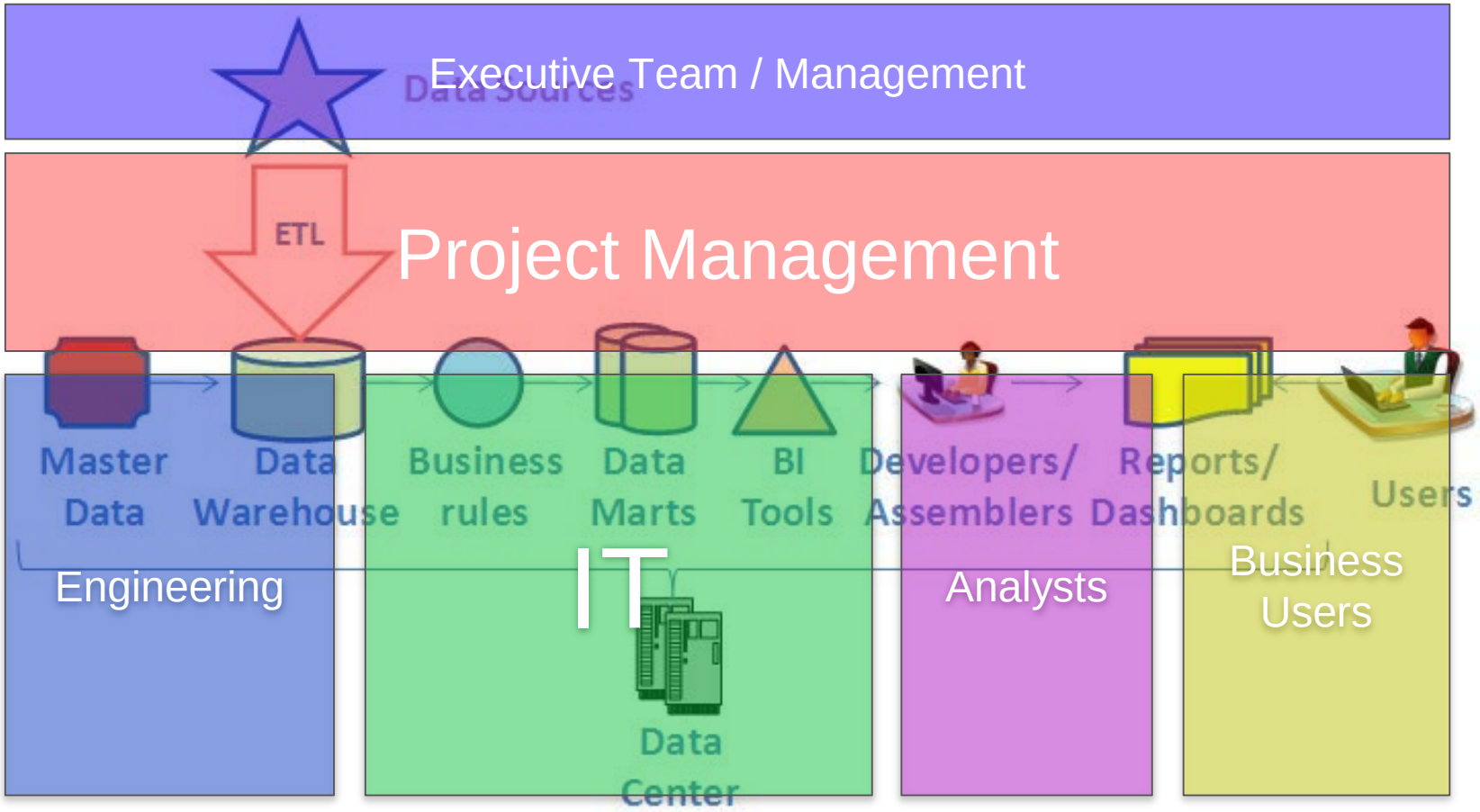
Teradata, HP Vertica, IBM Netezza, Oracle Exadata…



2006's Stack


Data Sources
ETL
Master Data
Data Warehouse
Business rules
Data Marts
BI Tools
Developers/ Assemblers
Reports/ Dashboards
Users
Data Center

2006's 5+ Tools & 6+ Teams



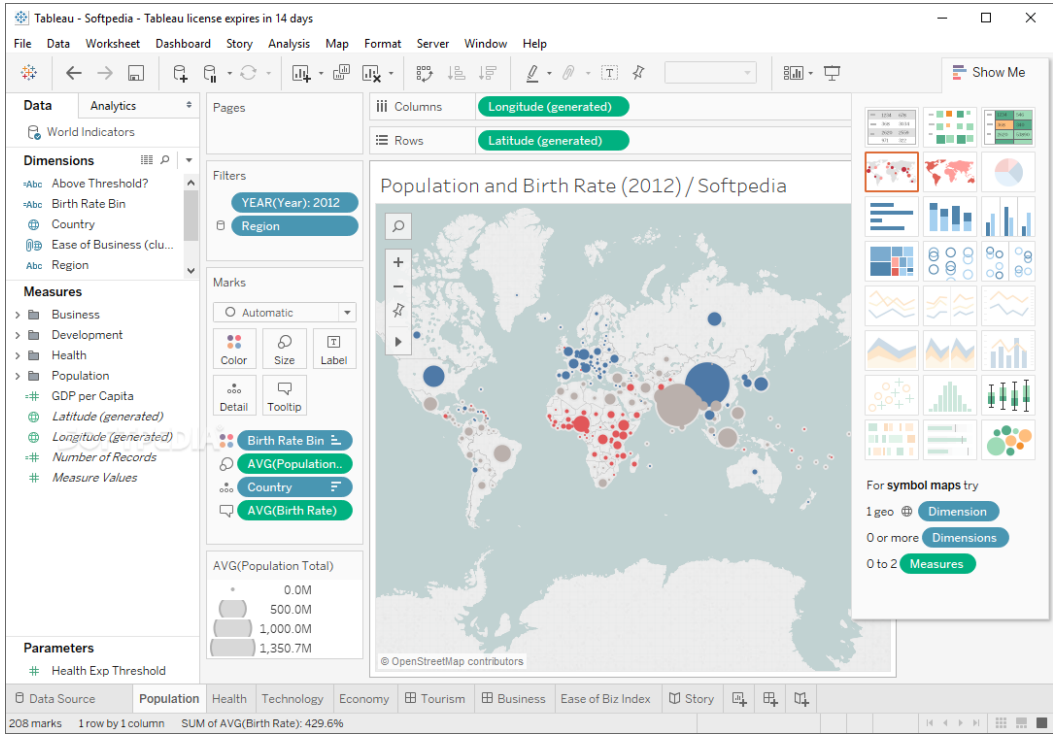
2008

2008's Self Service BI

Asking of data, why did this happen?

Tableau, Qlik

- Drill down
- Explore
- Still in data silos
- Multiple versions of truth



2008’s Data Governance

More data, more consumers. More complex data. Multiple version of the same truth. Decentralized BI tools.

Herding Cats!

2011

2011's Challenges with on -prem MPP Column store warehouses

Variety of Data

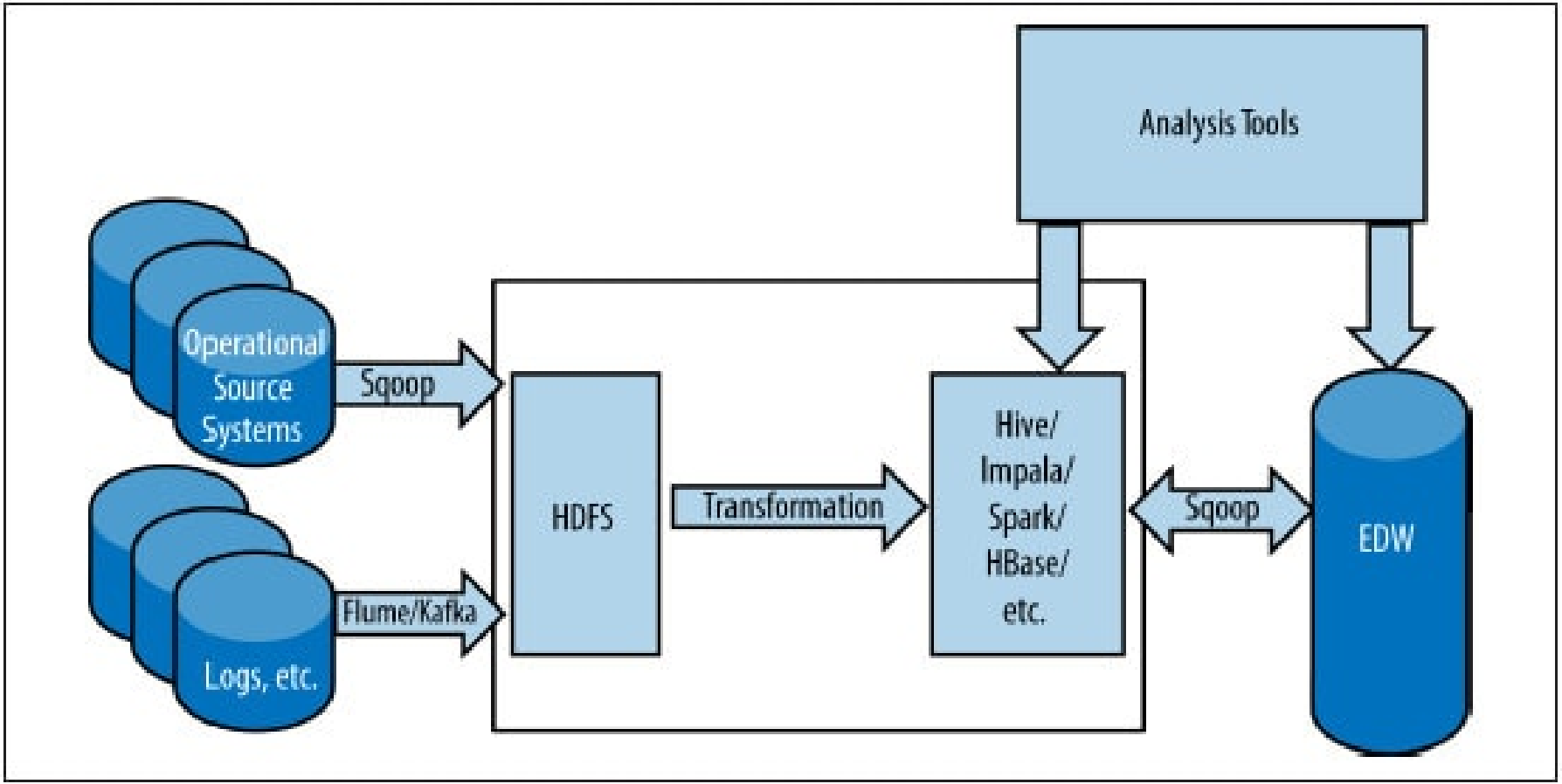
Variety of Analytics

2011's Hadoop to the Rescue!

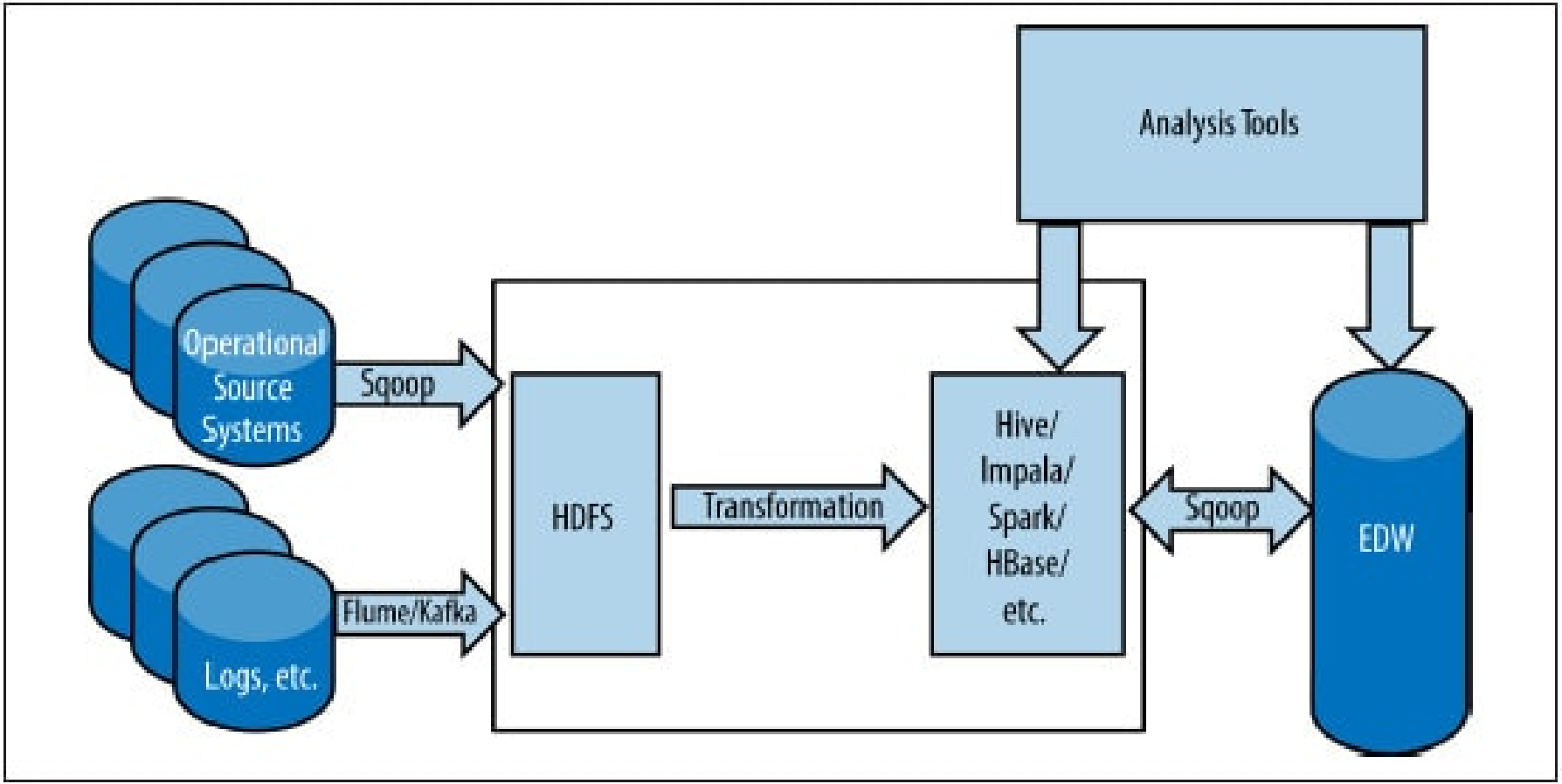
Built to:

- Scale to Big Data
- Handle all forms of data
- Allow any type of analytics

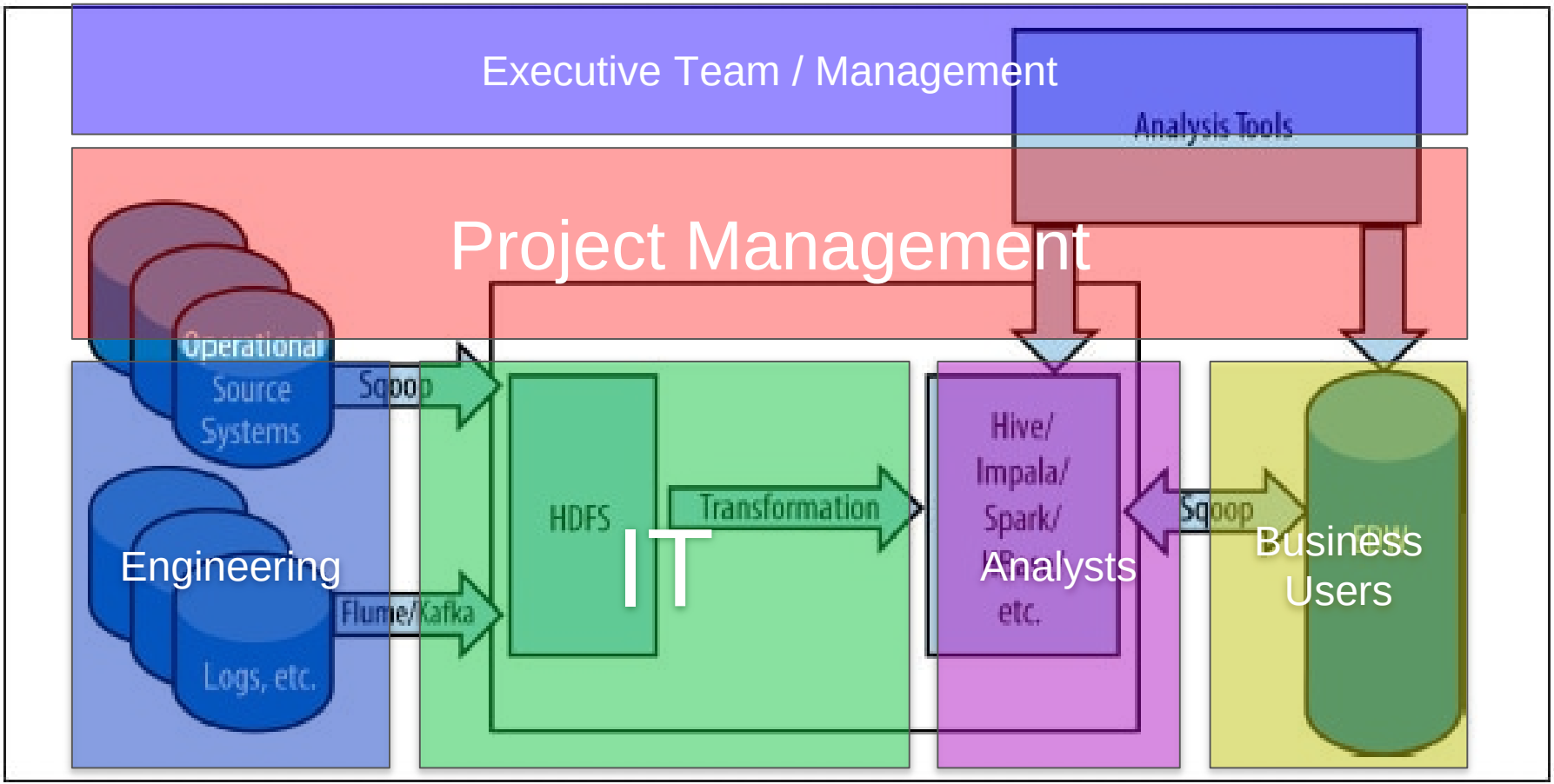
2011's Hadoop Stack



2011's Total Stack = 6+ Tools



2011's Still complicated team structure = 6+ Teams



2013

2013's Issues with Hadoop

Easy to dump data into a hadoop data lake…
hard to manage data and extract value.

- Complicated low level setup & maintenance
- Requires experienced development teams

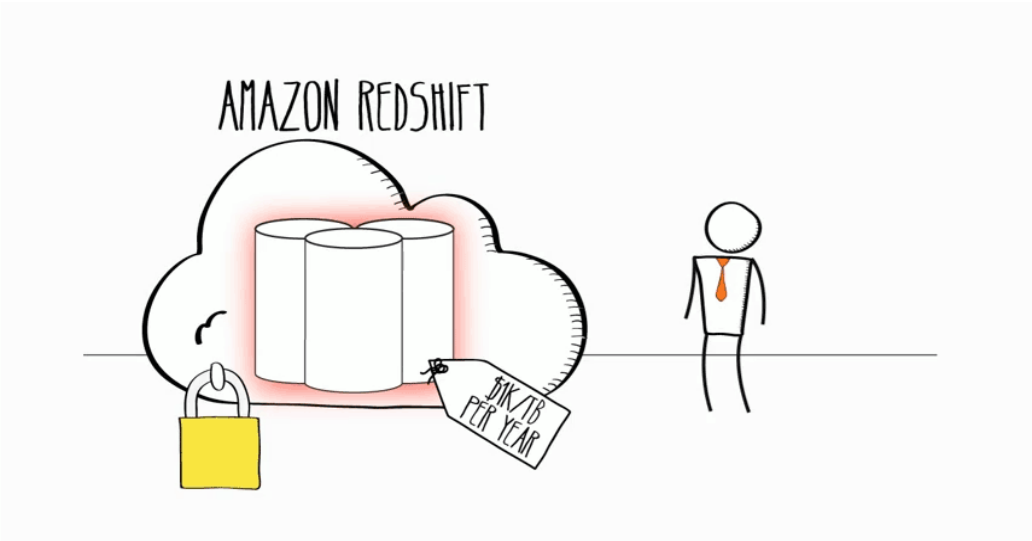
Ultimately companies end up sending data from Hadoop to SQL database for Analytics.

Dead end!

2013's - MPP Column Store in the Cloud - Redshift!

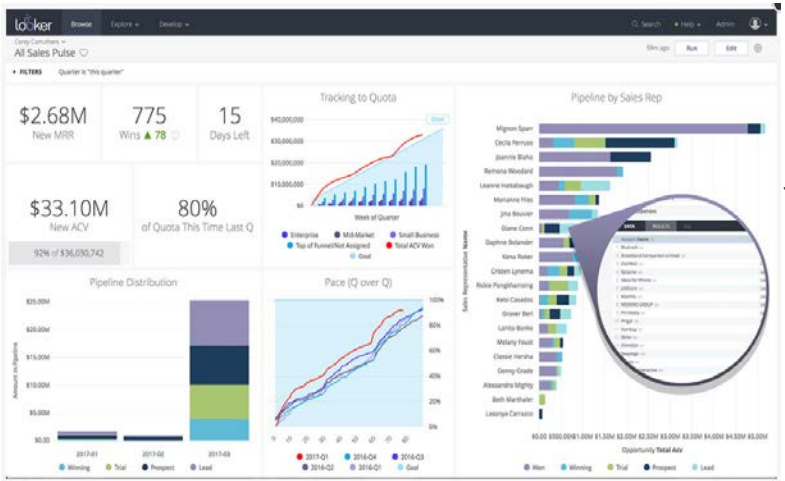
Fast, affordable EDW on AWS - awesome!

- MPP Scales
- Far less expensive than on-prem Column Store EDW
- Fairly easy to resize clusters etc
- 1 GB of data $0.05



2013 Cloud -Native Self Serve BI

Goal: Allow both centralized control of data, but also self serve to entire company.



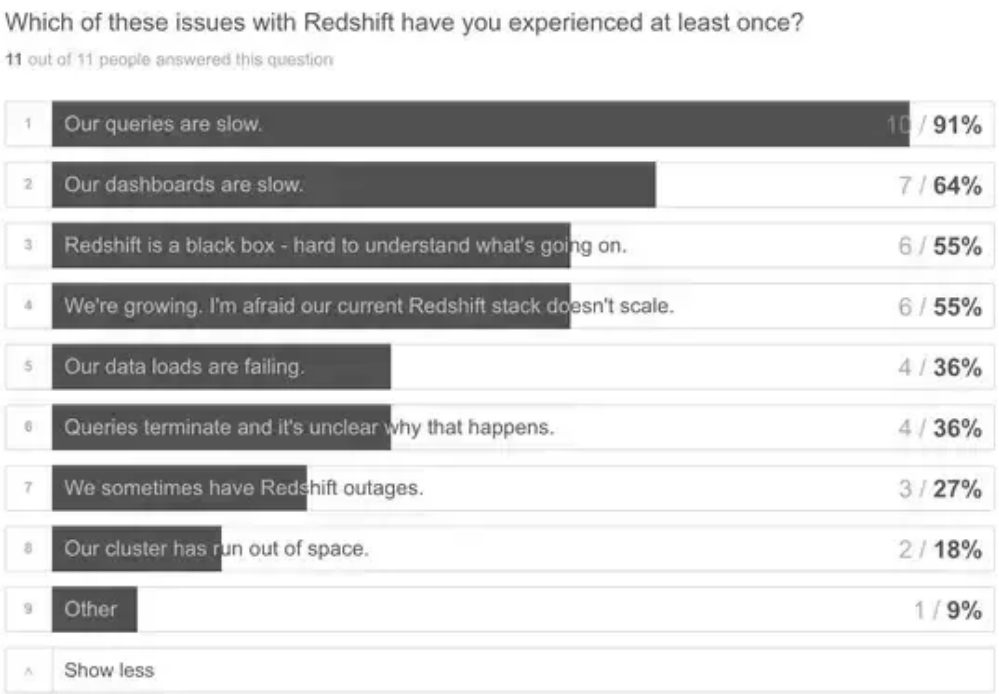
Looker

Make data so accessible, it starts to change the culture at the company to be more data driven.

- Using data to try to predict future
- Single version of the truth
- Full data accessibility
- Super fast, query directly against the DWH

2015

2015's Challenges with Redshift



"There's a 99% chance that the default configuration will not work for you!"

~ Lars Kamp

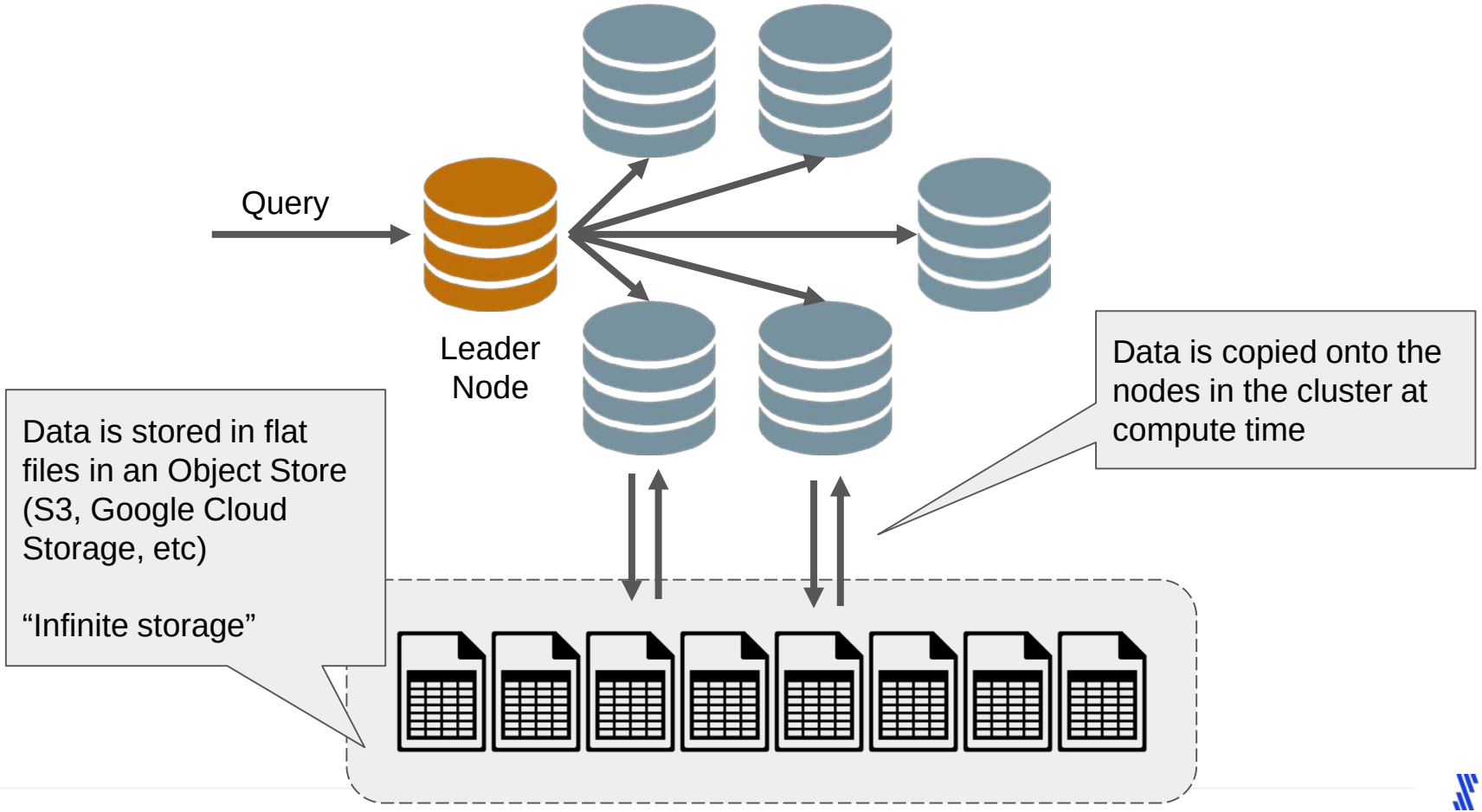
https://www.intermix.io/ - https://www.quora.com/What -problems -have-you-faced -while-working -with -Amazon -Redshift

Fivetran

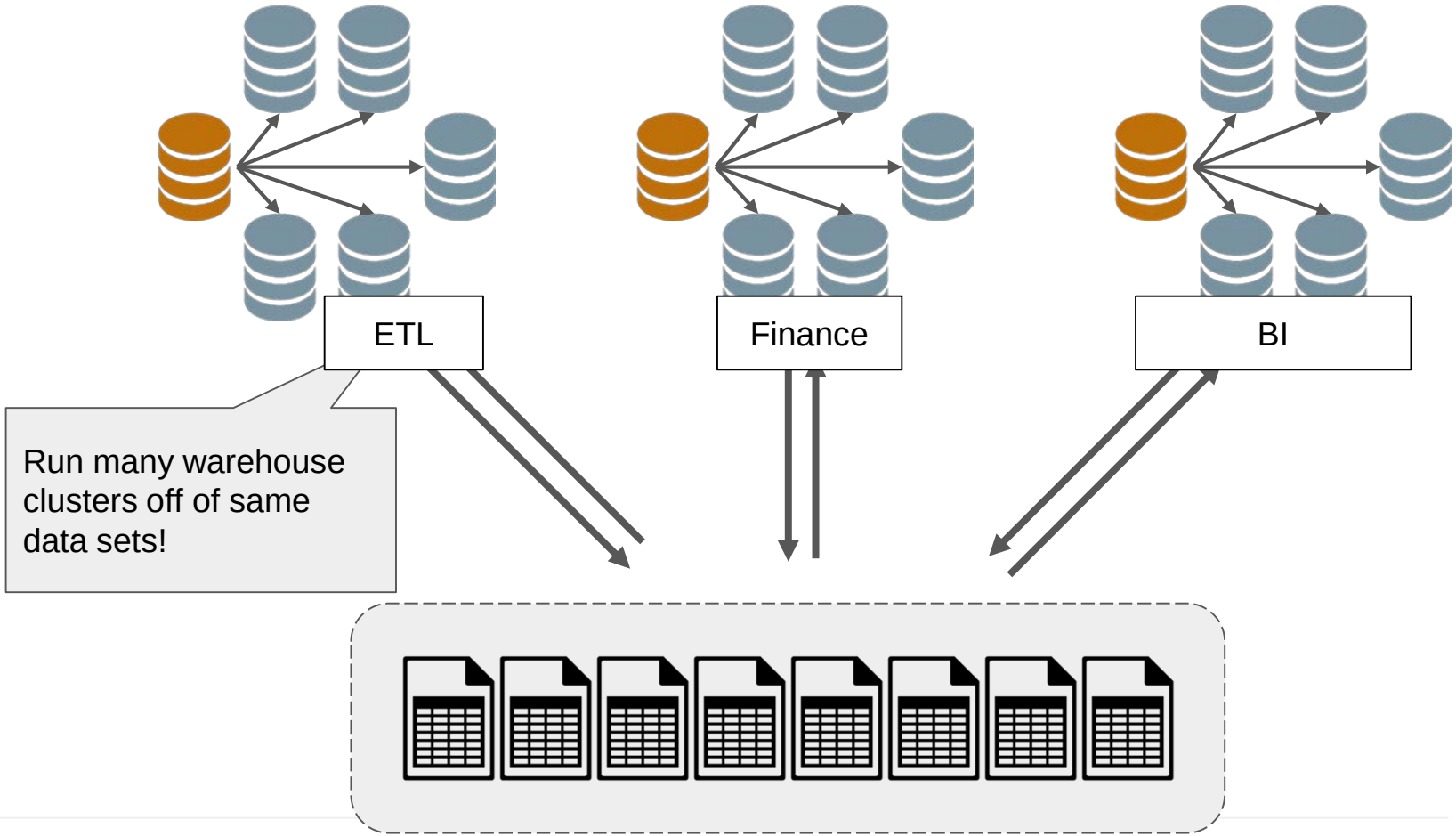
2015s Cloud -Native Column -store MPP Data Warehouses

1. Separation of compute & storage
2. Zero infrastructure management
3. Structured & Unstructured data
4. Instantly Scalable Compute

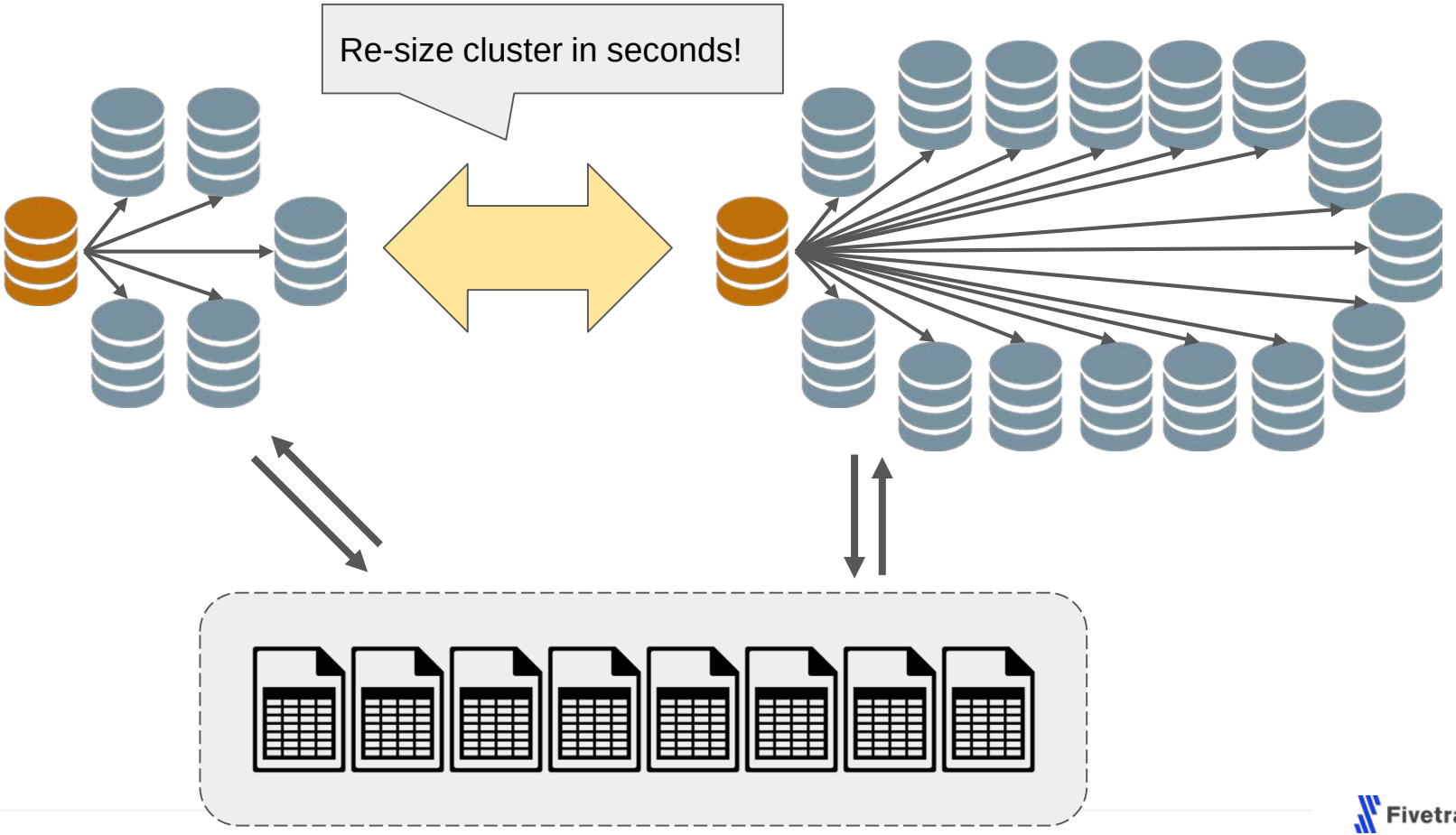
Separation of Compute & Storage



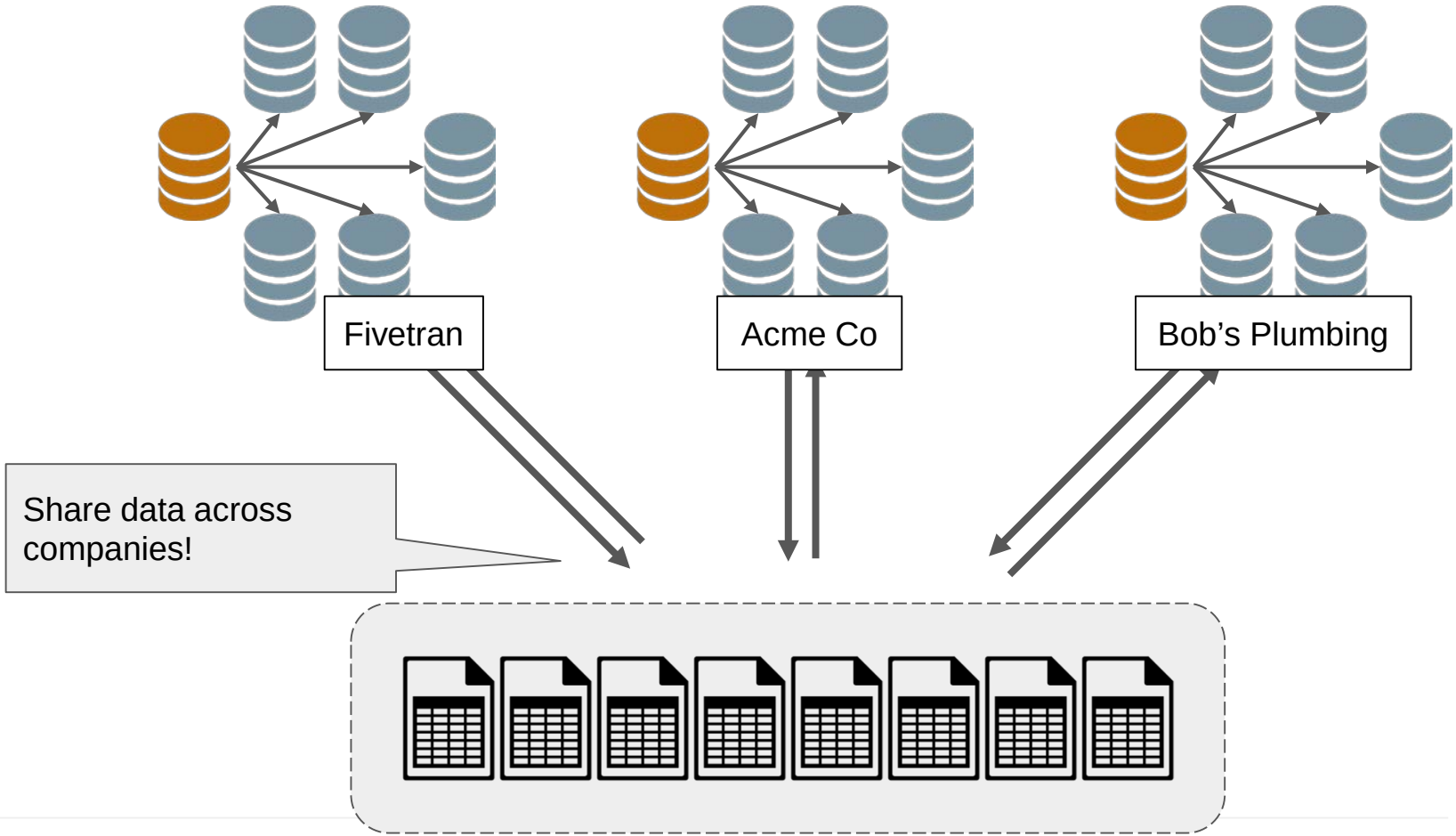
No more queue issues!



Elastic Compute



Data Sharing



How does this affect ETL?

Recap of changes

Warehouses

2000 OLAP

2006 On-prem Column Store MMP

2011 Hadoop

2000 Cloud Column Store MMP

2015 Cloud Native Column Store MMP

BI

2000 Monolithic Rigid BI

2008 Self Serve BI

2013 Centralized Cloud Native Self Serve BI

ETL

2000 Custom ETL

??

Challenges with ETL from 2000's

ETL was optimized for slow on -premise OLAP data warehouses, with massive storage constraints.

Optimized for pulling from on -premise enterprise applications

Extensive Setup

Ongoing Maintenance

Extensive
Planning

2015's Shift in company structures

With move to cloud, IT teams are shrinking.

Analyst at the front of self serve BI and want:

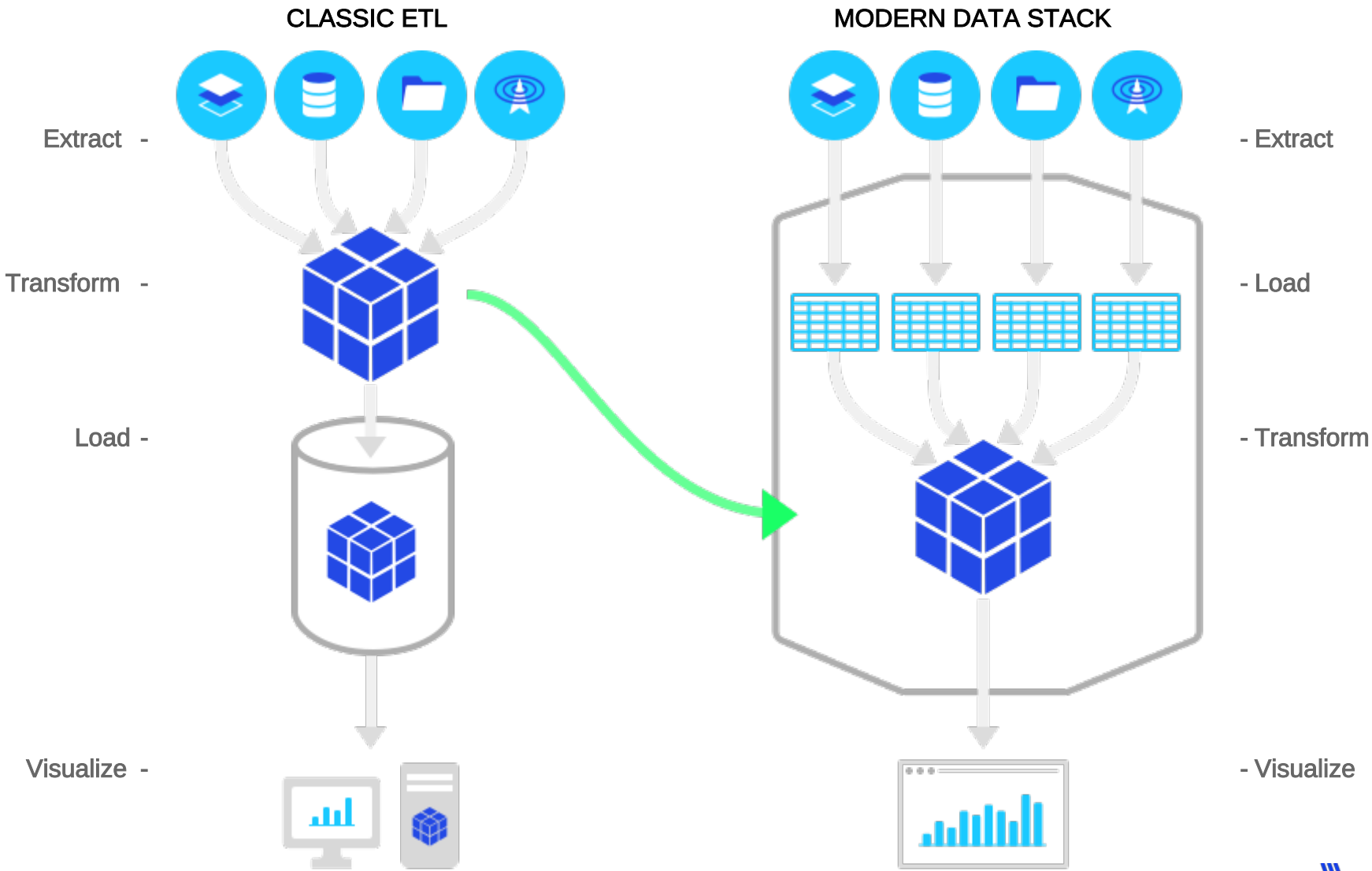
- Simple Infrastructure
- fully managed services
- wholistic control over stack

Other changes since 2000

Rise of cloud applications

Drop in data storage 1GB = $0.02

Agile workflows


CLASSIC ETL
MODERN DATA STACK
Extract -
Transform -
Load -
Visualize -
- Extract
- Load
- Transform
- Visualize

Fivetran

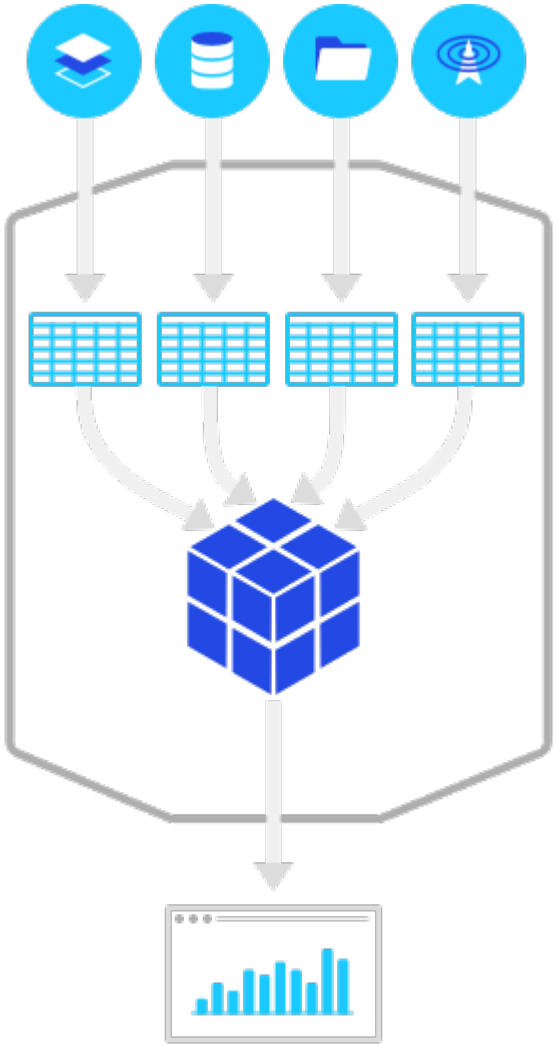
Agile Cloud -Native Self Serve Analytics - MODERN DATA STACK

FULL SCHEMA REPLICATION (DATA LAKE)

STANDARDIZED SCHEMAS

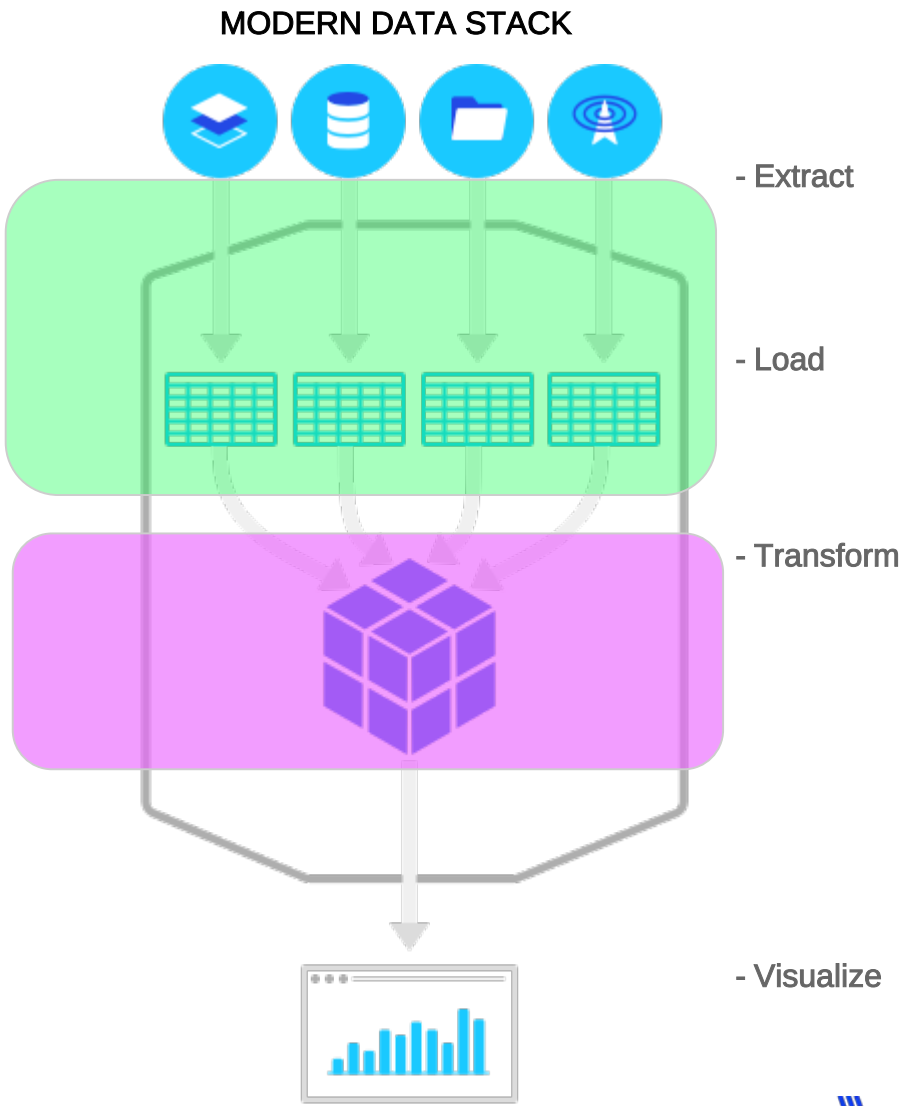
SQL-BASED MODELING & TRANSFORMATIONS

CENTRALIZED & COLLABORATIVE

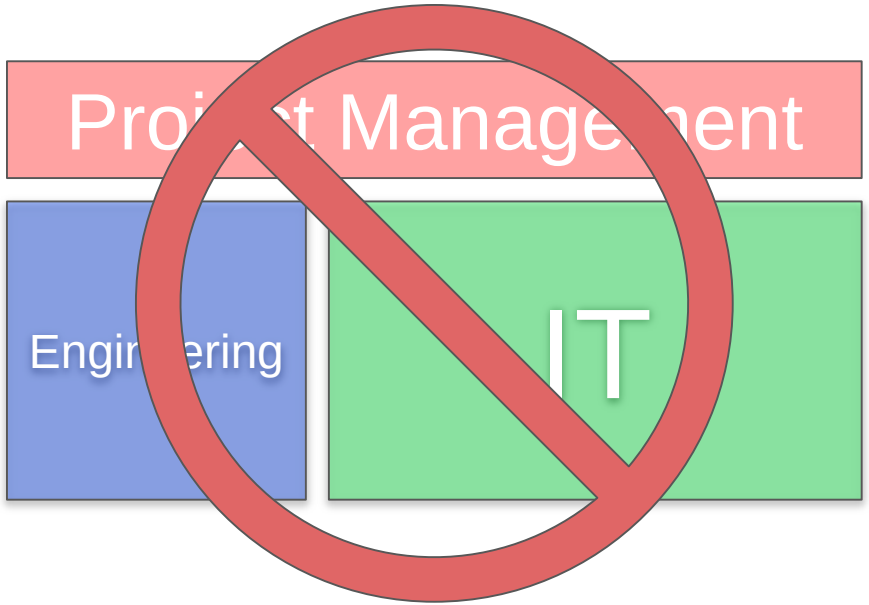


- Extract
- Load
- Transform
- Visualize

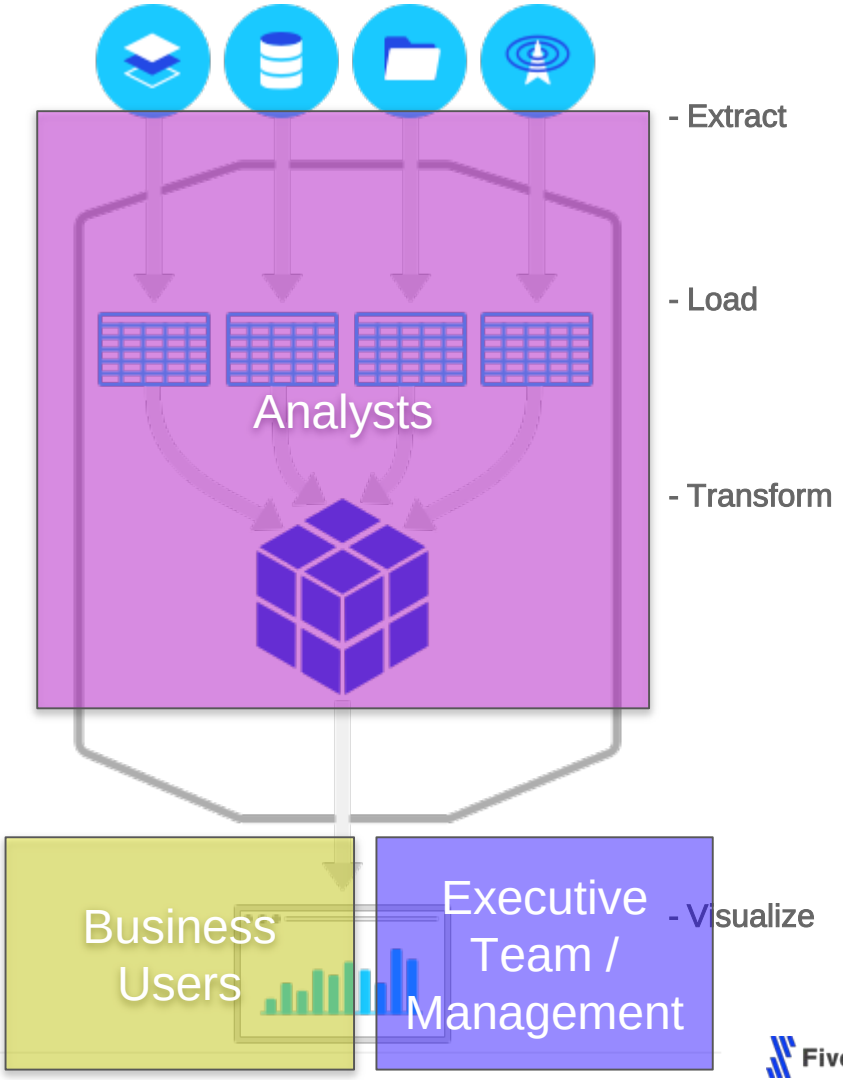
Modularize replication (Extraction & Load) from Transformation (Data gov)



Simplifies your management stack 3+ Teams



MODERN DATA STACK



Fivetran

Recap of changes

Warehouses

2000 OLAP

2006 On-prem Column Store MMP

2011 Hadoop

2000 Cloud Column Store MMP

2015 Cloud Native Column Store MMP

BI

2000 Monolithic Rigid BI

2008 Self Serve BI

2013 Centralized Cloud Native Self Serve BI

ETL

2000 Custom ETL

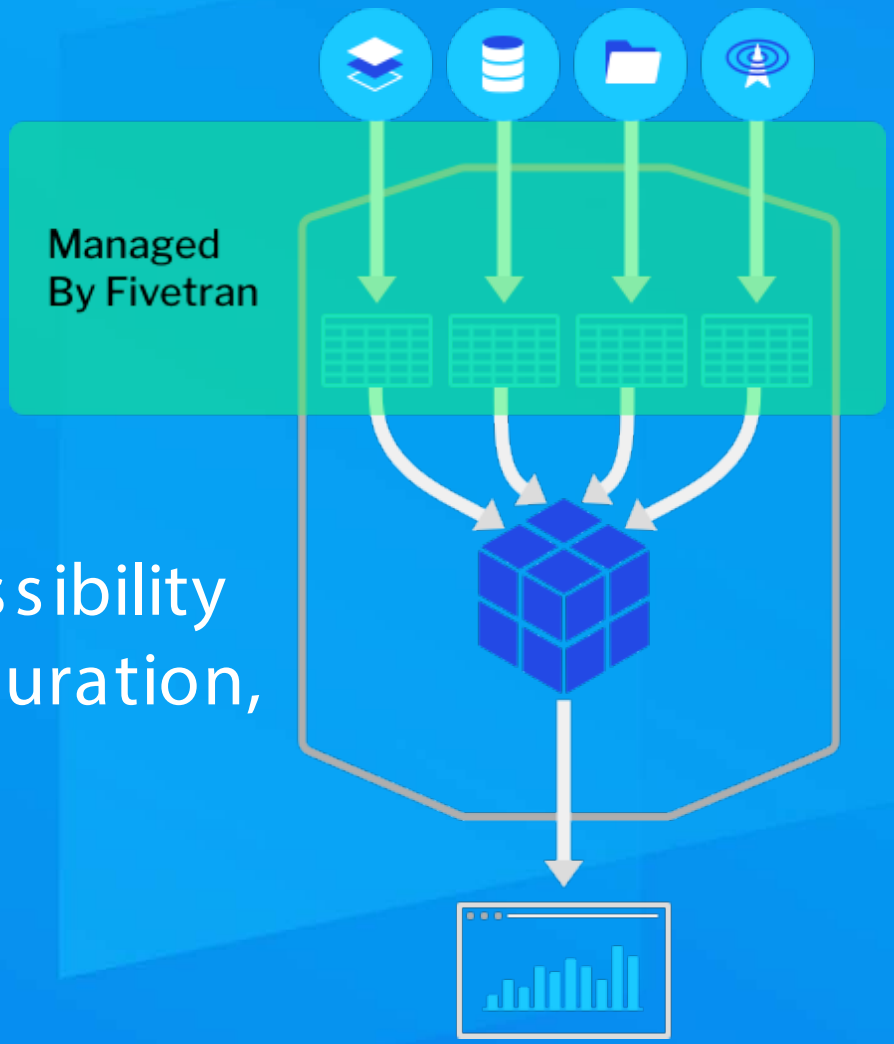
2015 ELT Separate EL & Transform

Fivetran

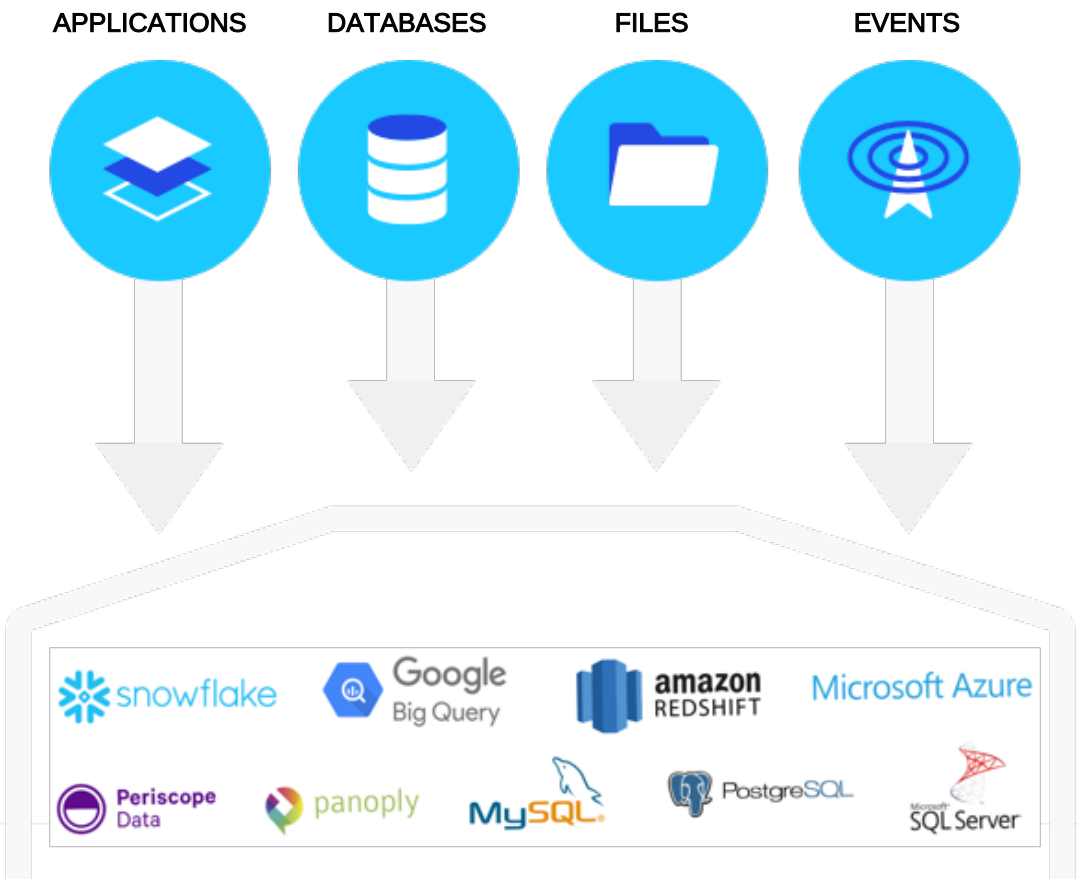
Fivetran

Zero Configuration, Zero Maintenance, Data Pipelines

Fivetran helps you achieve data accessibility with its zero configuration, zero maintenance data pipelines



Data pipeline as a service



Your warehouse

Applications

Apple Search Ads
Asana
AdRoll
Bing Ads
Braintree Payments
Desk.com
DoubleClick
Dynamics (365, GP, AX)
Eloqua
Facebook Ad Insights
Freshdesk
Front
Github
Google Adwords
Google Analytics
Google Play
Help Scout
HubSpot
Hybris
Instagram
Intercom
iTunes
Jira
Klaviyo
LinkedIn Ads
Magento
MailChimp
Mandrill
Marketo
Mavenlink
Mixpanel
NetSuite SuiteAnalytics
Optimizely
Pardot
Pinterest Ads
QuickBooks Online
ReCharge
Recurly
Sailthru
Salesforce
SalesforceIQ
SAP Business One
SendGrid
Shopify
Stripe
SugarCRM
Twitter Ads
Xero
Yahoo Gemini
Zendesk
Zendesk Chat (Zopim)
Zuora

Databases

Amazon Aurora
Amazon RDS
Azure SQL Database
DynamoDB
Google Cloud SQL
Heroku
MariaDB
MongoDB
MySQL
Oracle DB
PostgreSQL
SQL Server

Files

Amazon Cloudfront
Amazon Kinesis Firehose
Amazon S3
Azure Blob Storage
CSV Upload
Dropbox
Email CSV Ingester
FTP
FTPS
Google Cloud Storage
Google Sheets
SFTP

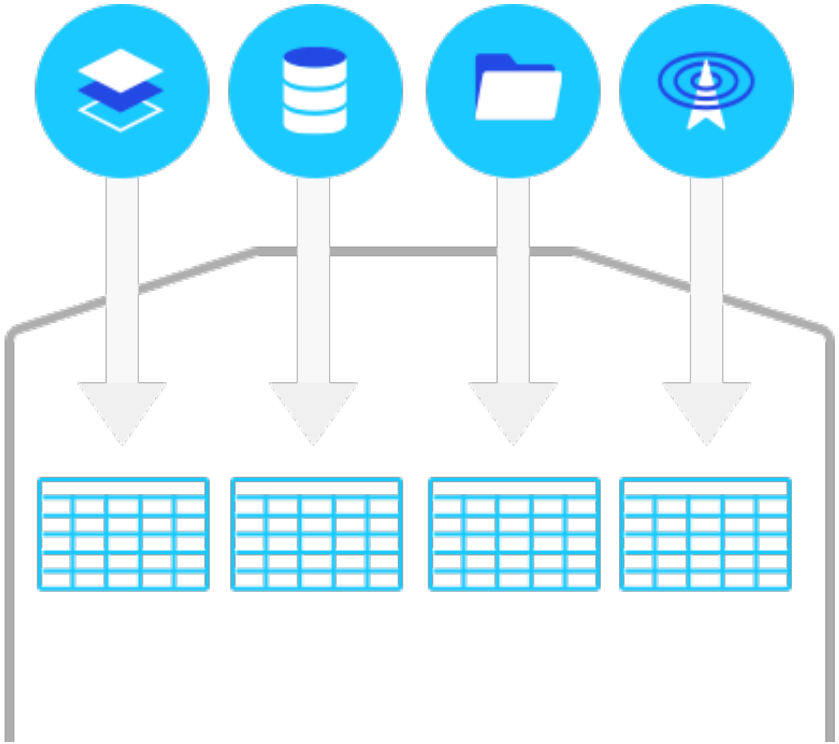
Events

Google Analytics 360
Segment
Snowplow
Webhooks

For an updated list of data sources visit fivetran.com/directory

Fivetran

Authenticate, and we do the rest...



1. Pull historical Data
2. Normalize
3. Create Schema/Tables & Load Data

Update

Fivetran Data Normalization Behavior

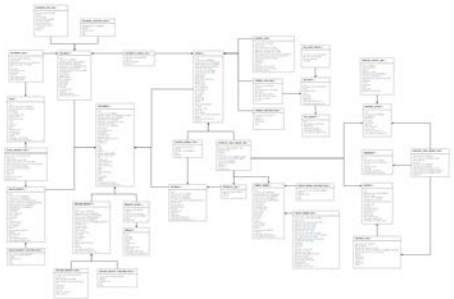
SOURCE

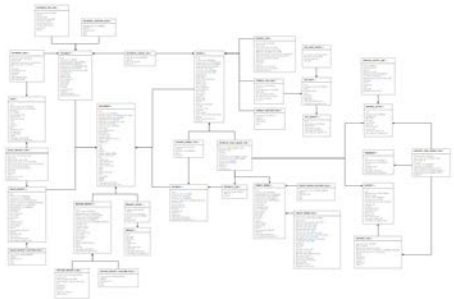
WAREHOUSE

We normalize Denormalized Data
(from APIs)

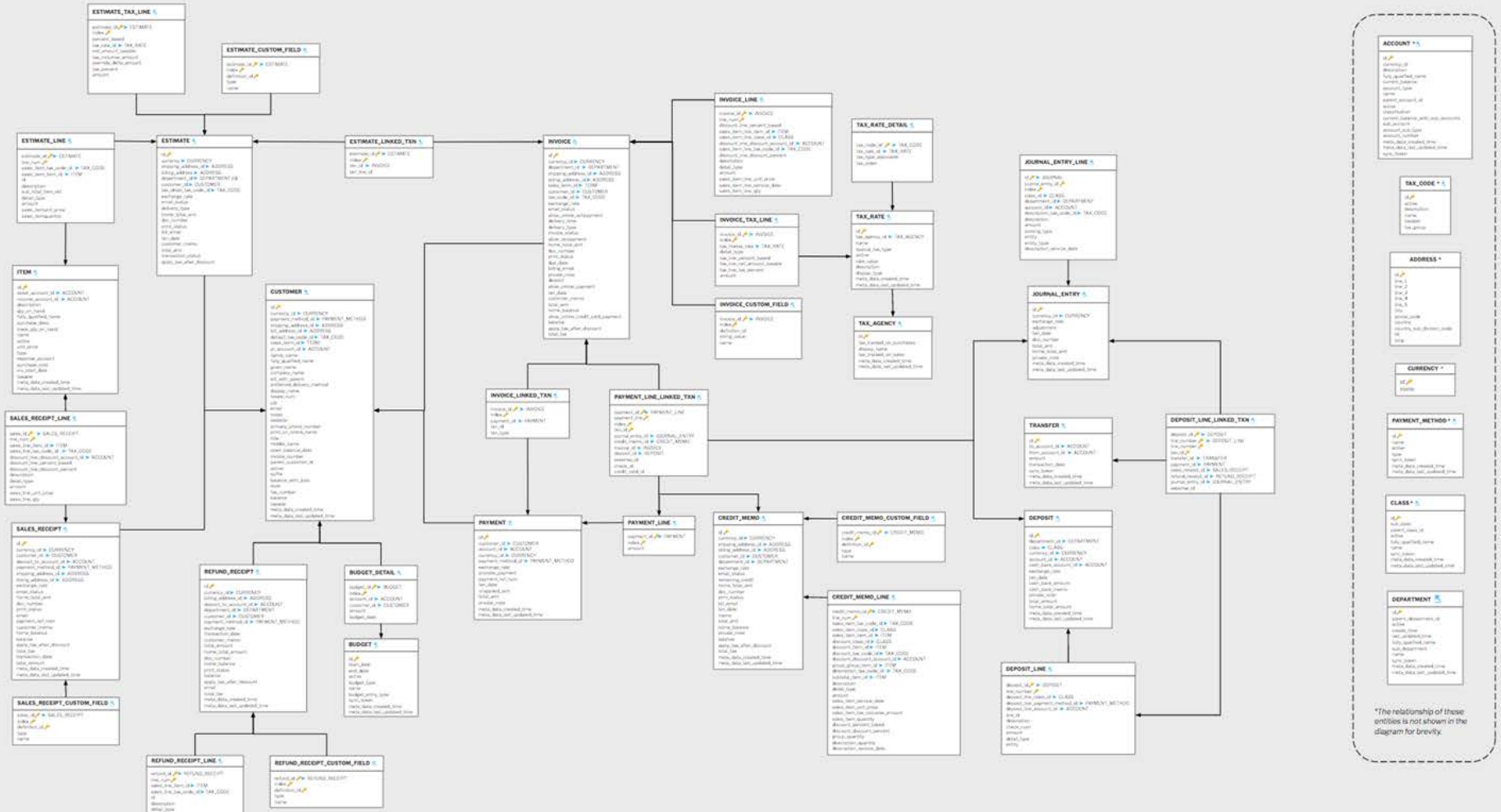
```
{menu: {
  id: "file",
  value: "File",
  popup: {
    menuitem: [
      {value: "New", onclick: "CreateNewDoc()"},
      {value: "Open", onclick: "OpenDoc()"},
      {value: "Close", onclick: "CloseDoc()"}
    ]
  }
}}
```

We replicate Normalized Schemas
(Databases, SFDC, Netsuite)



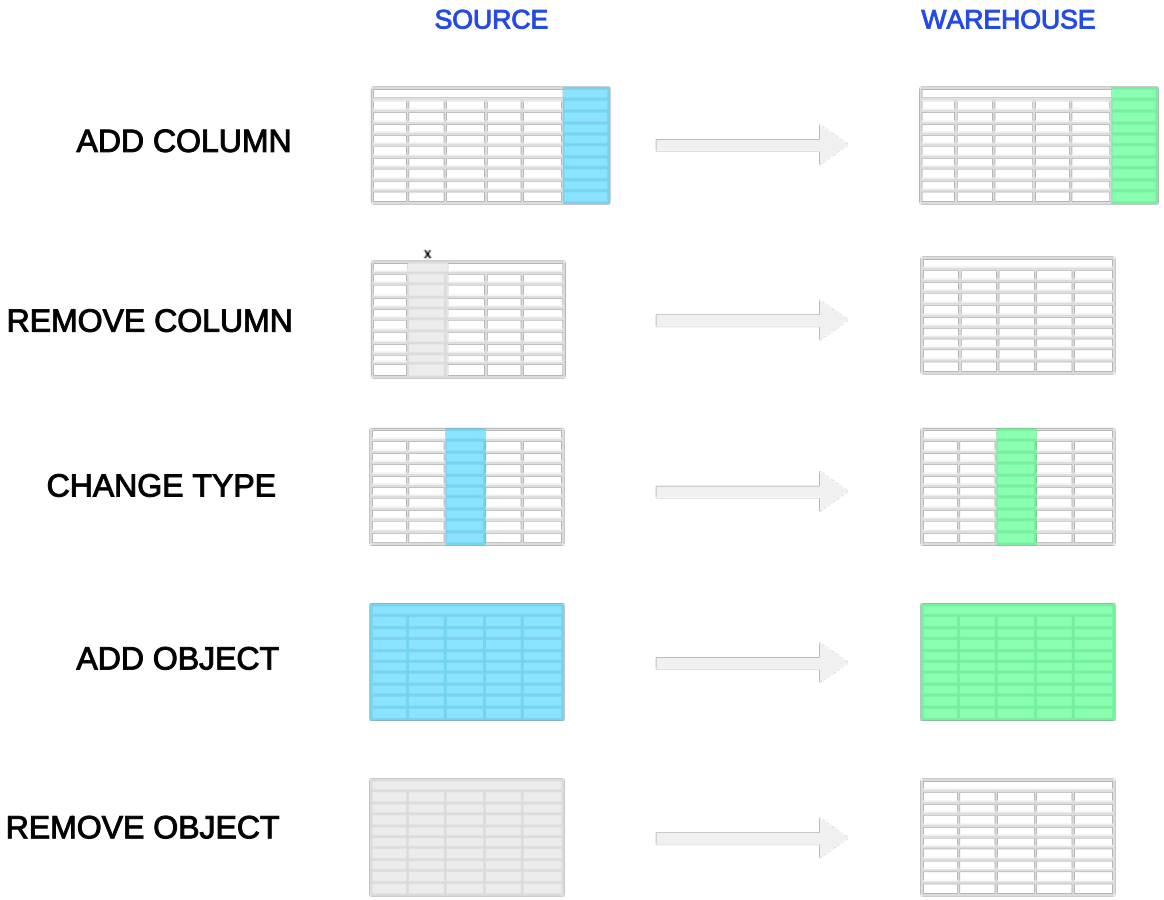


Standard Schemas - ERDs



Incremental Batch Updates

Automatic Schema Migrations



Complexity compounds. Automate, standardize and simplify as much of your stack as you can.

Recap of changes

Warehouses

2000 OLAP

2006 On-prem Column Store MMP

2011 Hadoop

2000 Cloud Column Store MMP

2015 Cloud Native Column Store MMP

BI

2000 Monolithic Rigid BI

2008 Self Serve BI

2013 Centralized Cloud Native Self Serve BI

ETL

2000 Custom ETL

2015 ELT Separate EL & Transform

Fivetran

What’s coming next?

Feedback, questions, thoughts?

Taylor@fivetran.com

Fivetran

Zero Configuration, Zero Maintenance, Data Pipelines