data iku

DATA SCIENCE SALON
MIAMI 2018

Stop Wasting Time – Case Studies in Production Machine Learning

Yashas Vaidya
Technical Lead, Alliances-US
yashas.vaidya@dataiku.com

2 X 2

Two questions, two case studies

- When putting things into production ...
  - why do Data Scientists take too long?
  - why does DS takes too long?
- Two case studies getting around those issues
  - Speeding up ETL steps to create insights
  - Case: Starting with a template and making reusable analytics

Data Science in production

Why the focus?

- Business lines are often the sponsors and need to see ROI from projects
- Creates value across the organization and provide DS teams with recognition
- From strategic insights to augmented (or automated) decision making

Wait a minute …

Where is the world is Ken?

data iku

Re-introduction

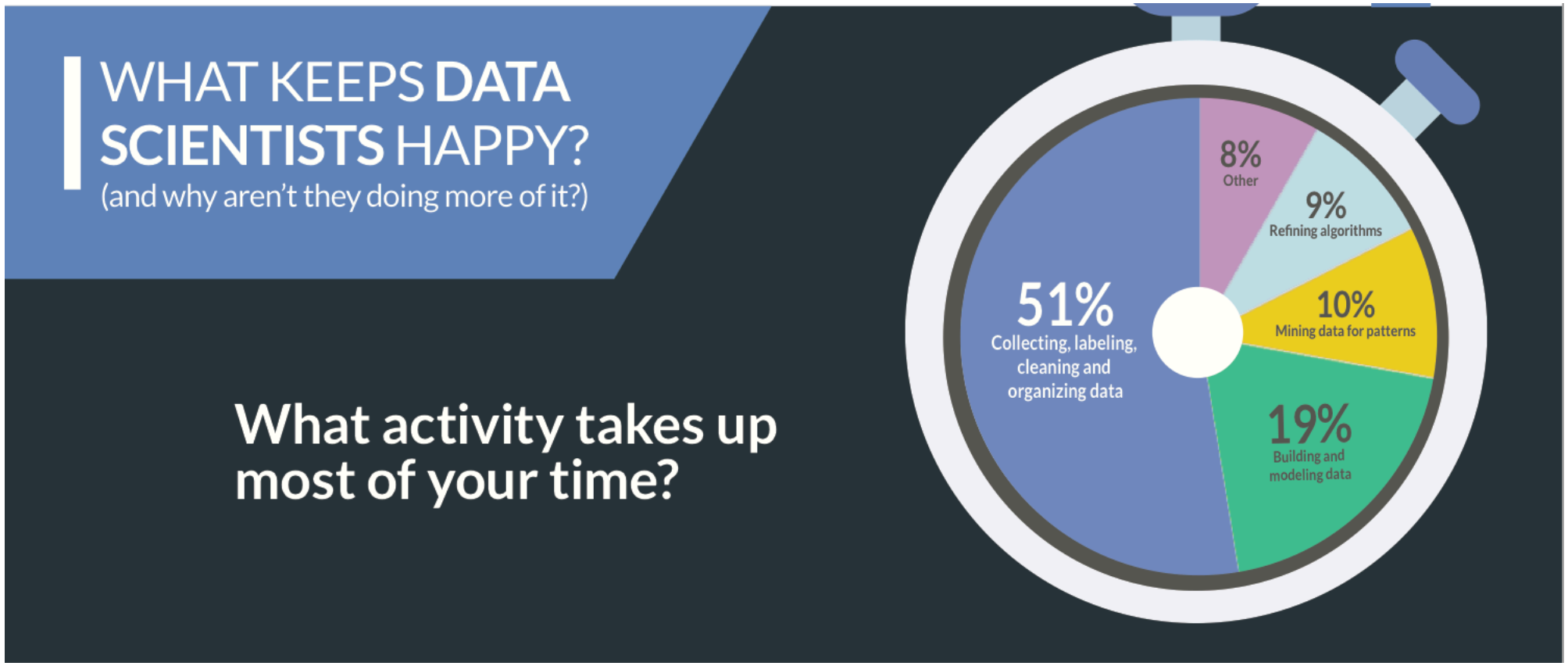
I'm Yashas and I am ...

- Evangelizer in the Alliances team,
- Training and empowering consulting and technical partners
- ABD ...

data iku

Data Scientists take too long

They mostly spend their time doing ETL



Source: 2017 Data Scientist Report
CrowdFlower: https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html

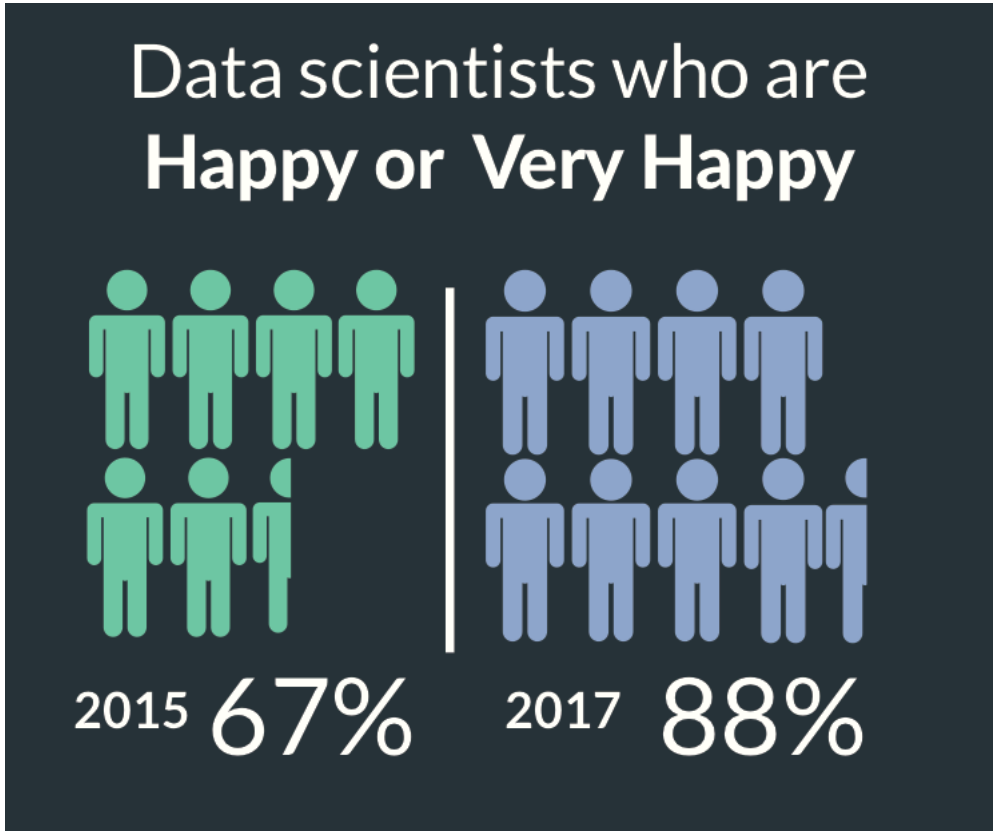
Why do Data Scientists take so long

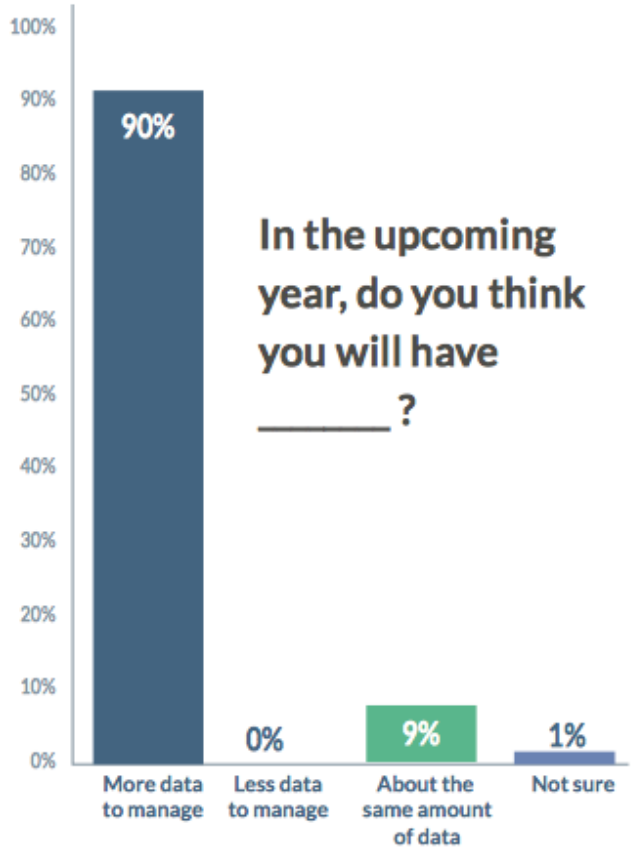
Several hypotheses

- H1: Data scientists have a hard time getting their hands on good quality data and thus must scrub away
- H2: Given their background, they are actually pretty good at cleaning and organizing data. So they focus on what they can do well and control.

Some evidence for both

Data scientists are usually happy data scientists, overwhelmed by data





Source: 2017 Data Scientist Report
CrowdFlower: https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html

Getting to production is hard to do

Why is data science hard to productionalize?

- different teams in charge
- lack of established platforms or frameworks
- overall low maturity on deployment skills and experience

McKinsey Institute: *« less than 10% of data science projects are deployed into production […] with average deployment times of 9 to 12 months »*

Case Study 1

First case study today

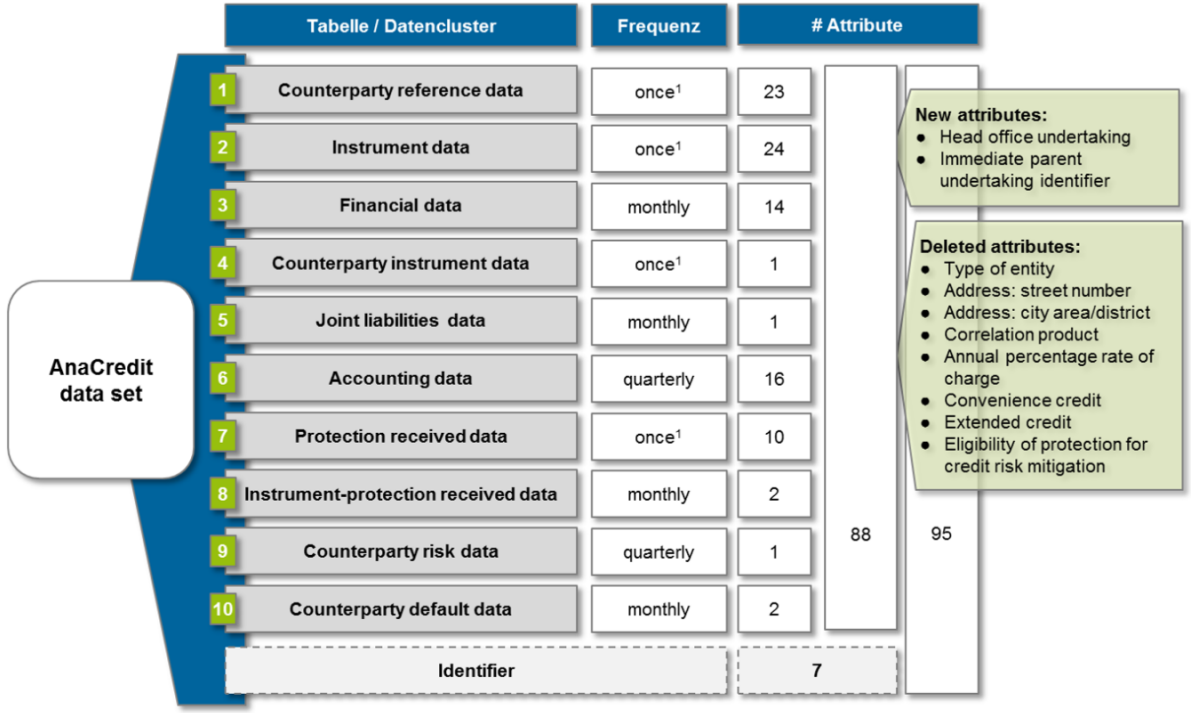
Regulatory Compliance

data iku

A bit of context

Anacredit regulatory reporting

- New European credit reporting regulation – AnaCredit
- Every financial institution must report a monthly dataset of loans over 25k with up to 95 attributes
  - instrument data
  - counterparty data
  - liability data…
- Progressive rollout in 2018 for 8 countries
- Large Banks and Insurance companies have provisioned around 10-15M € for the project



Source: ECB regulation on the collection of granular credit and credit risk data as of May 18th, 2016

→ How can we provide a innovative, differentiating solution to this problem ?

data iku

Building the Anacredit solution

Usage Scenarios and Approach

- Key Usage Scenarios
  - Build and automate financial reporting and in same format across 8 countries
  - Take into account corrections and quality tests, as well as rejections
  - Provide broader reporting on data quality and insights on portfolio management

- A best-of-breed approach to the project
  - Data Management & Predictive – Dataiku DSS
  - Regulatory watch and Risk Compliance – PwC
  - Self Service Analytics, Insights and Reporting – Tableau Software

Data Management

Focus on Agility and Repeatability

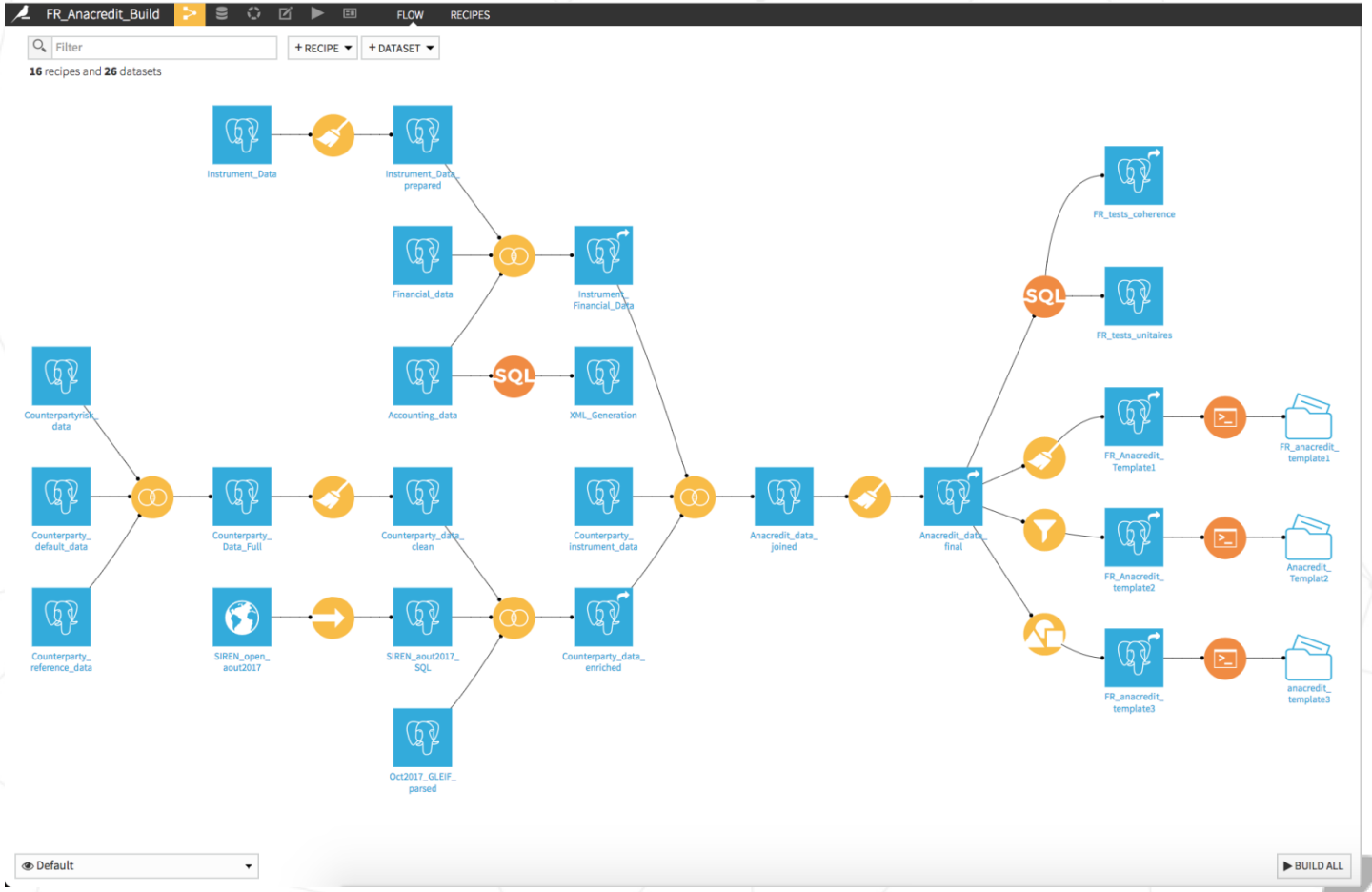
Data Processing Flow for Template 1

Working « prototype » of the dataset and templates within 2 months

- 2 Data Engineers
- 1 Business Analyst
- 1 Risk Analyst

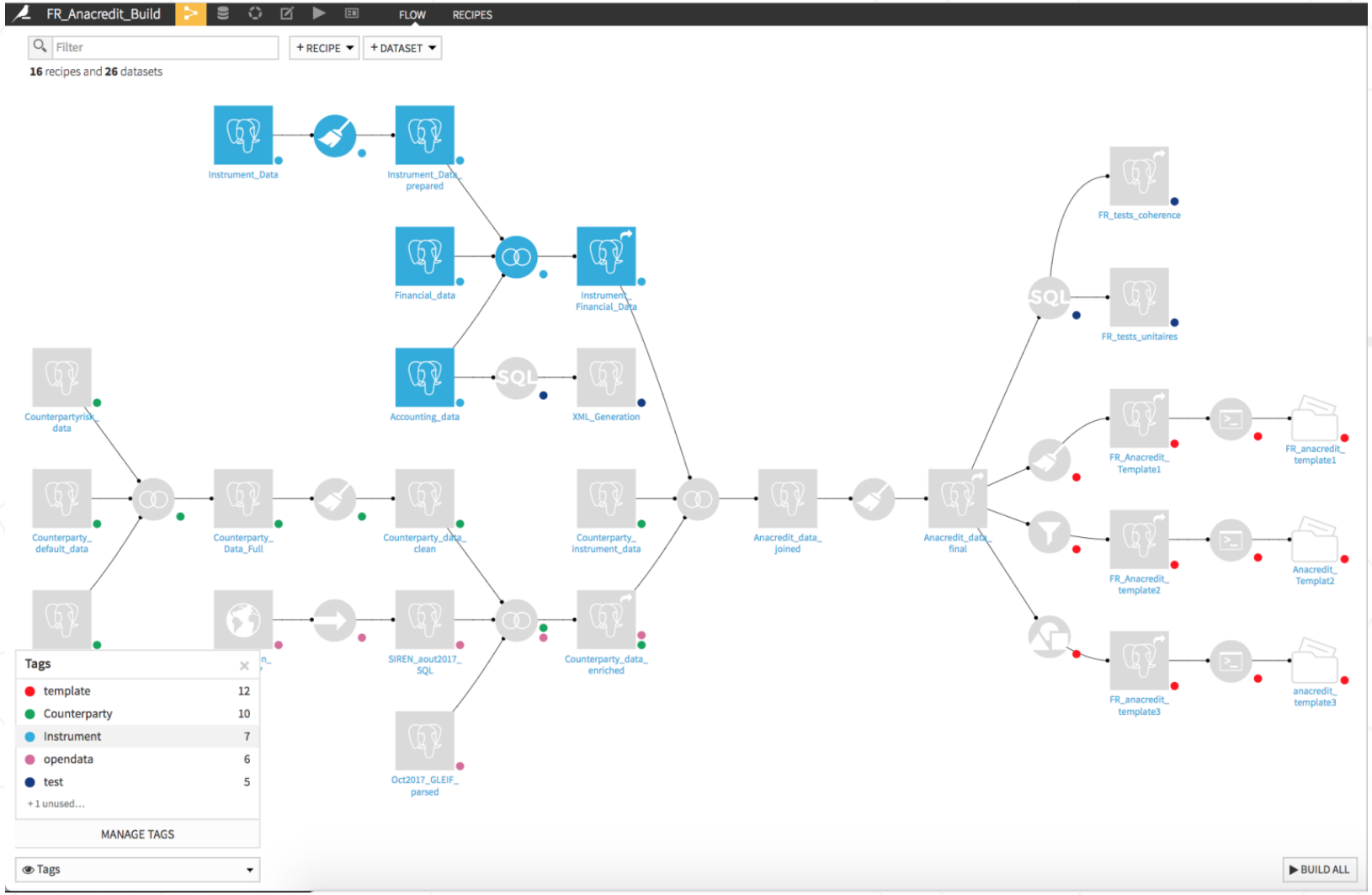
- Agile, documented, re-usable, runs « at scale » in the company's IS

- Articulated with ETL and DI tools



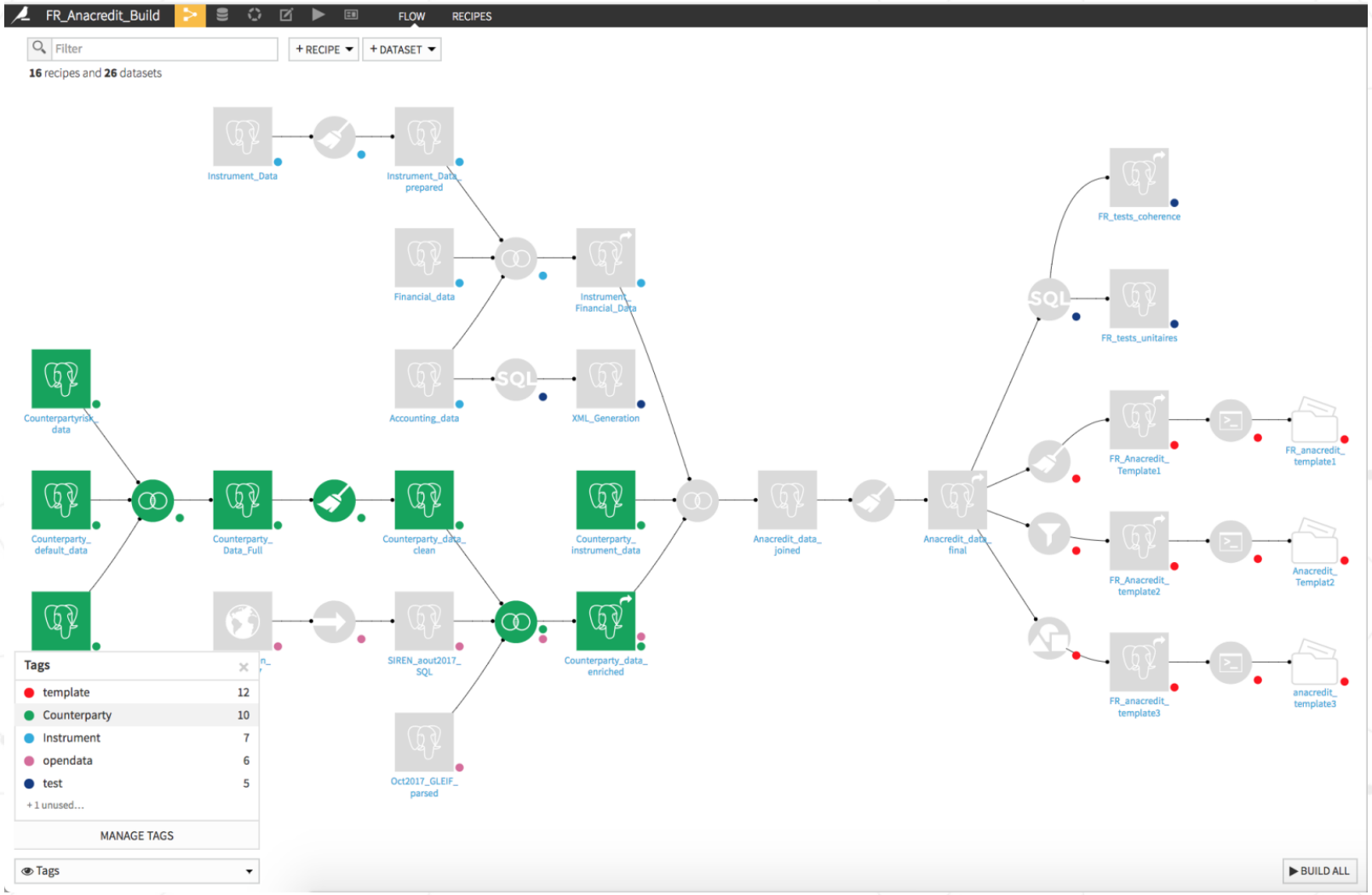
Data Management

Instrument Data



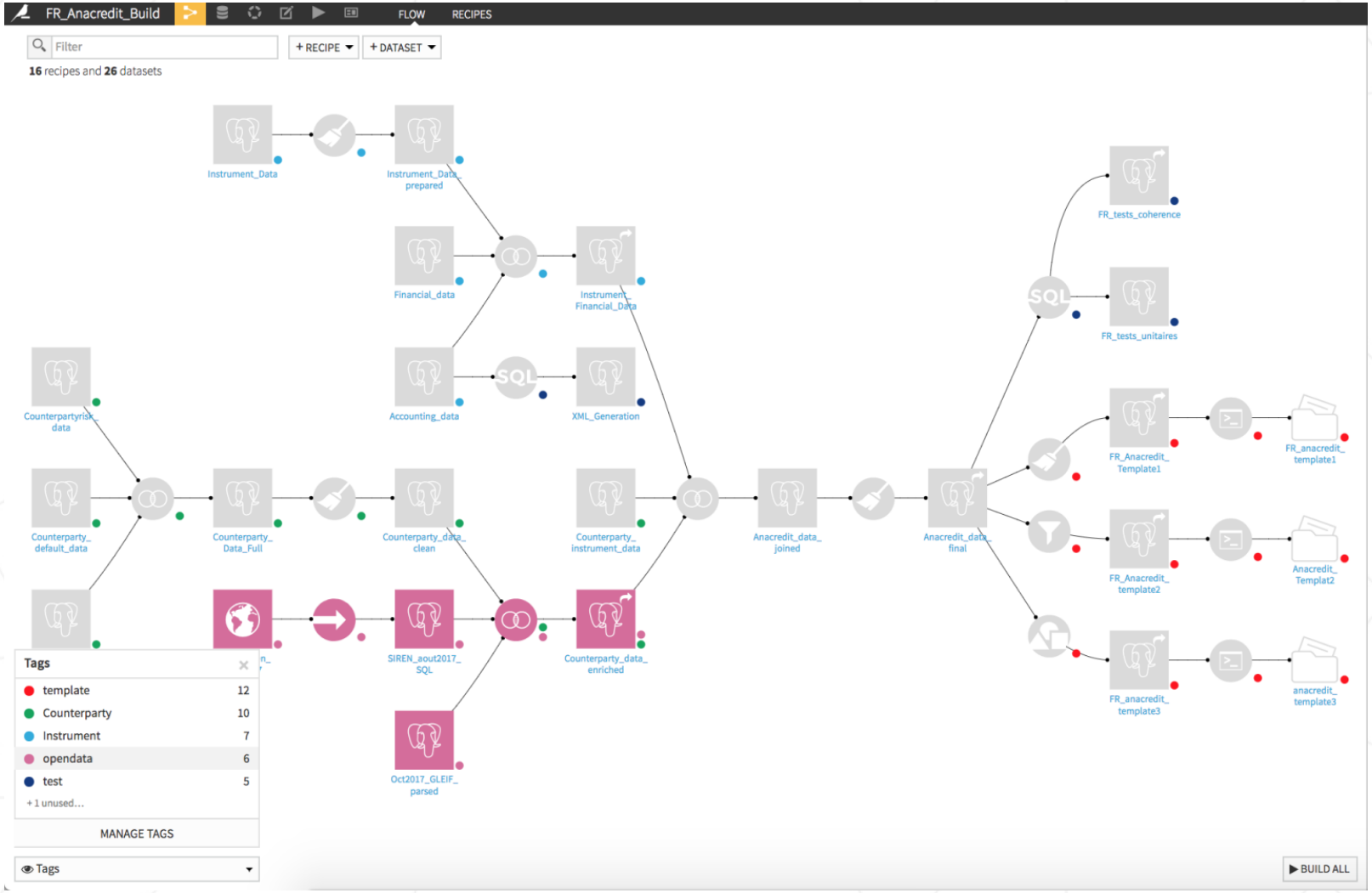
Data Management

Counterparty Data



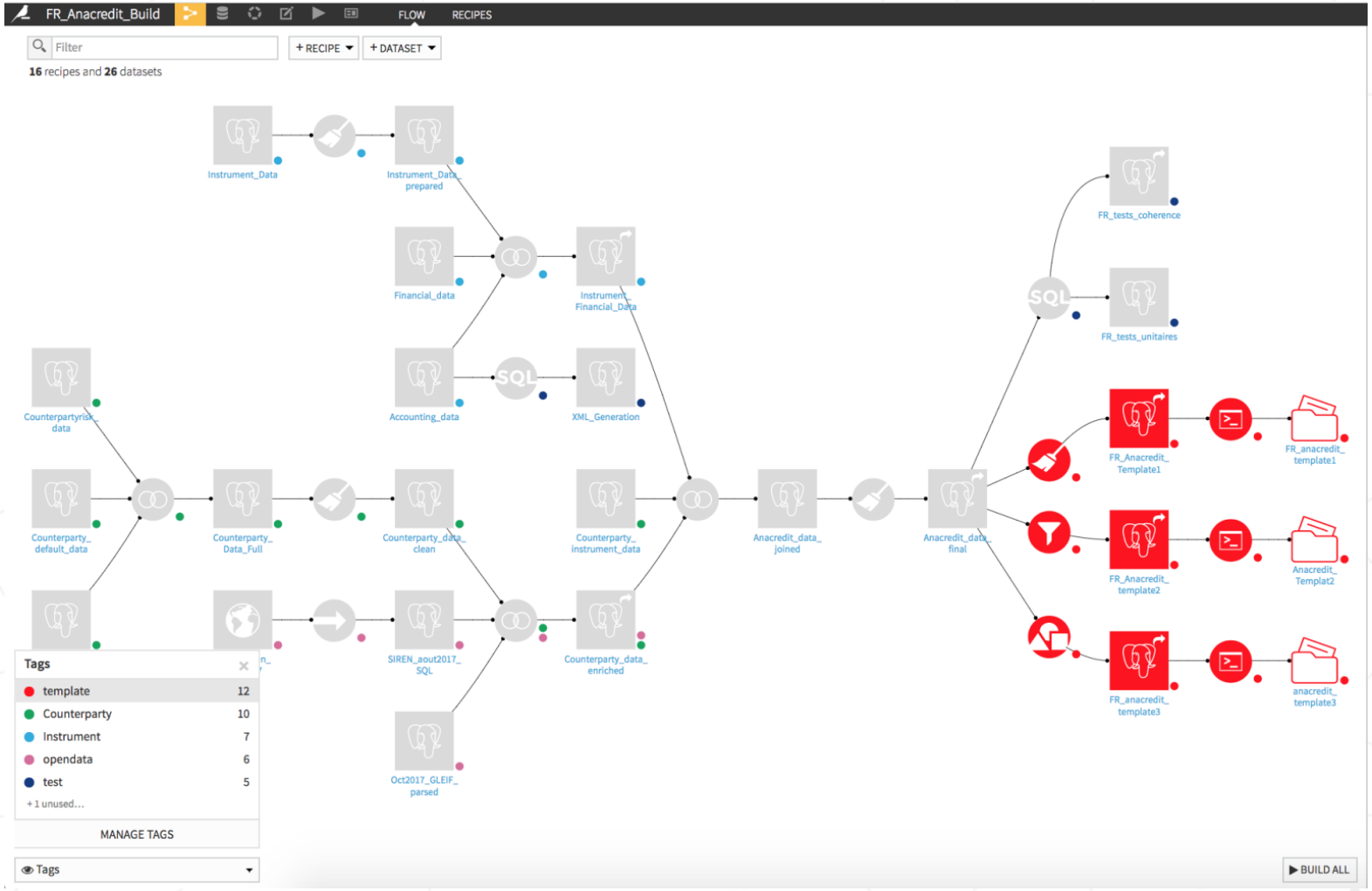
Data Management

Open Data



Data Management

Processing and Delivevering XML Templates

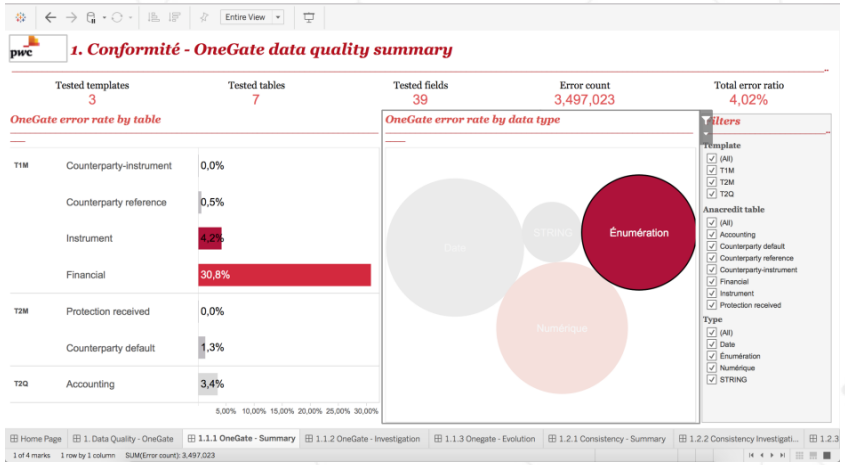


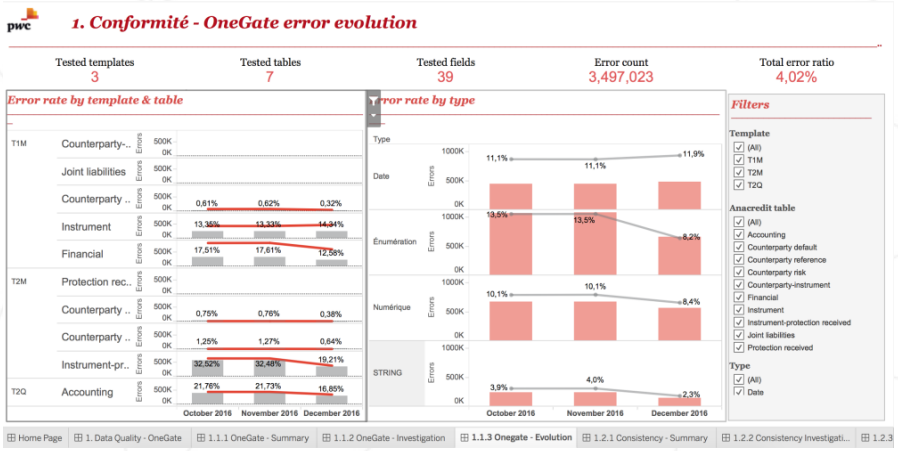
data iku

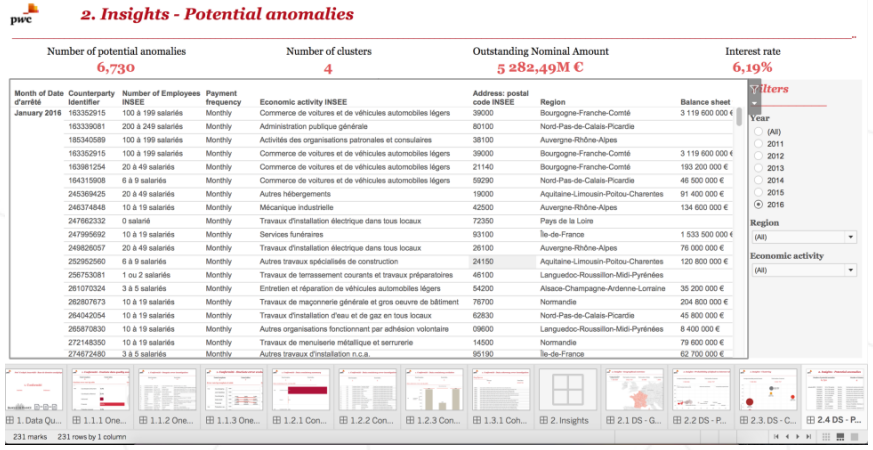
Analytics and Reporting

End user dashboards, improve analytics culture

- Data Quality and Compliance
- Evolutions over time of errors and rejects
- Risk portfolio reporting
- Predictive insights





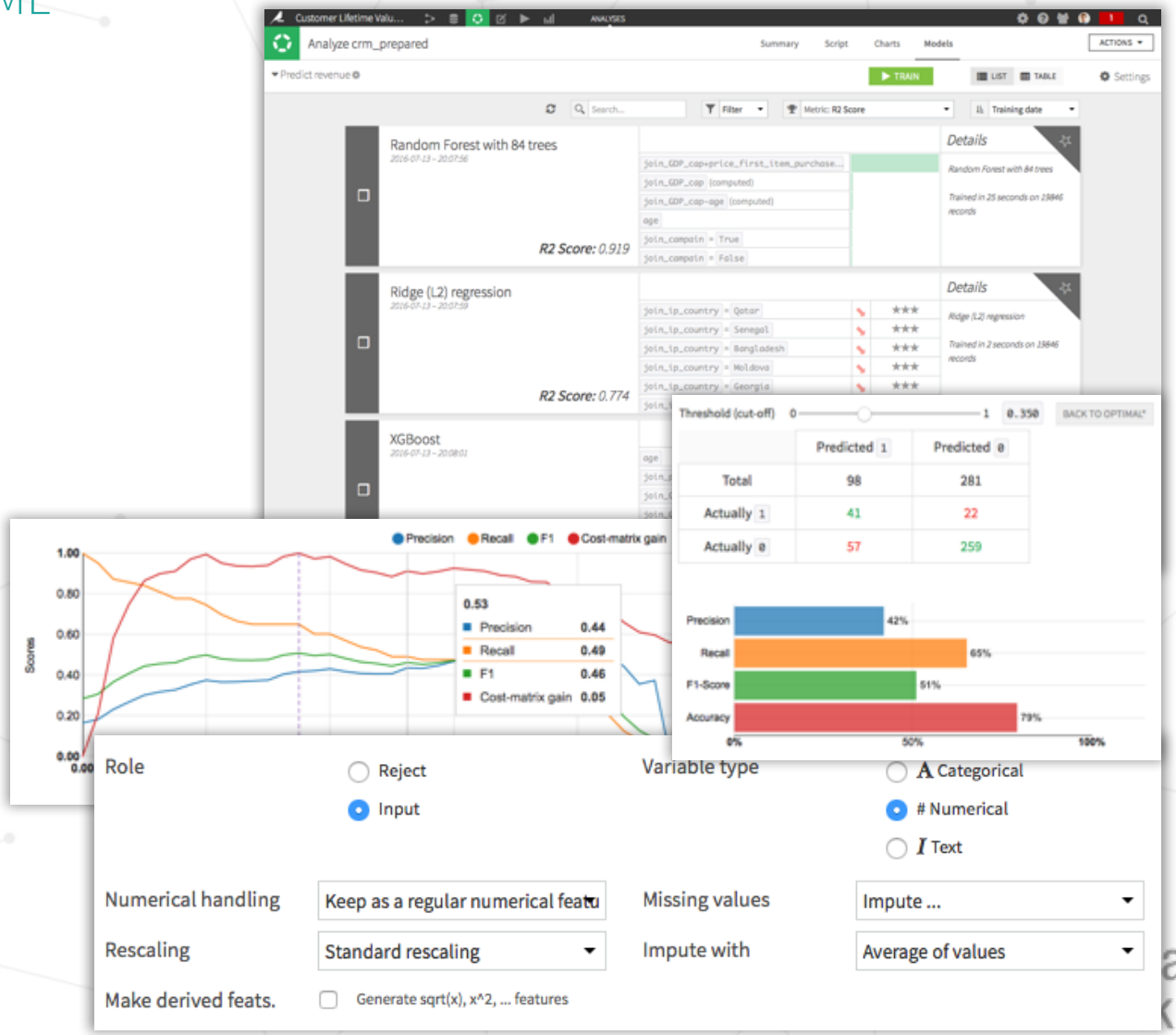


Machine Learning

Anacredit dataset is a great playground for ML

Fast exploration of multiple use cases to build business cases

- Automatic enrichment of data through fuzzy matching of LEI records
- Predicting outcome for late payments across loan fifecycle
- Anomaly detection on loans (Isolation Forest)



Case Study 2

Real-time recommendation system at scale

A production conundrum

- Power a sales platform used by 4,000 different clients

- First project
  - Provide customized recommendations (3 models per client)
  - Recommendations should available in real time (API endpoint)

Quote #170808-0001 - Atlas, Inc.

Quotation | Additional Information | Customer Information | Documents

Date Created
01/01/2016

Status
Preparing

Comment
A comment

Date Modified
01/01/2016

Market
United States (USD)

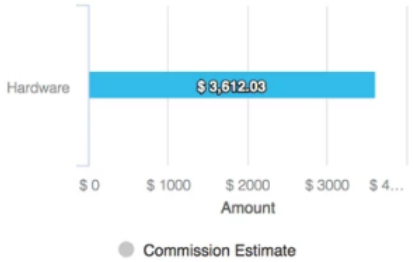
Revision
1

Products

Manage Items

Items	Quantity	List Price	Extended List Price	Discount Percent	Discount Amount	Net Price	Ext
T-NID 1.0 Network Interface Device	100	$500.00	$50,000.00	22	($11,000.00)	$390.00	$3
WRT 5.1 Wireless Router	50	$350.00	$17,500.00	28	($4,900.00)	$252.00	$1

Commission Estimate



Estimated Total Earnings
$3,612.03

Totals

Total List Price	
Discount	
Total Net Price	
Subtotal	
Shipping	Weight: 17 lbs
Taxes	00.00%

Grand Total

Thunderbridge AI Mandates

Your Priority Actions

New (11) | Closed (0) | Deferred (2)

Recommendation

Network Manager is popular among customers with similar cart contents.

$ 10,000.00 per user, annually retail
$ 9,150.00 with 8.5% discount

Add

Recommendation

Web-Managed Switches is more common among similar customers. Consider recommending as of add-on to the T-NID 1.0 Network Interface Device.

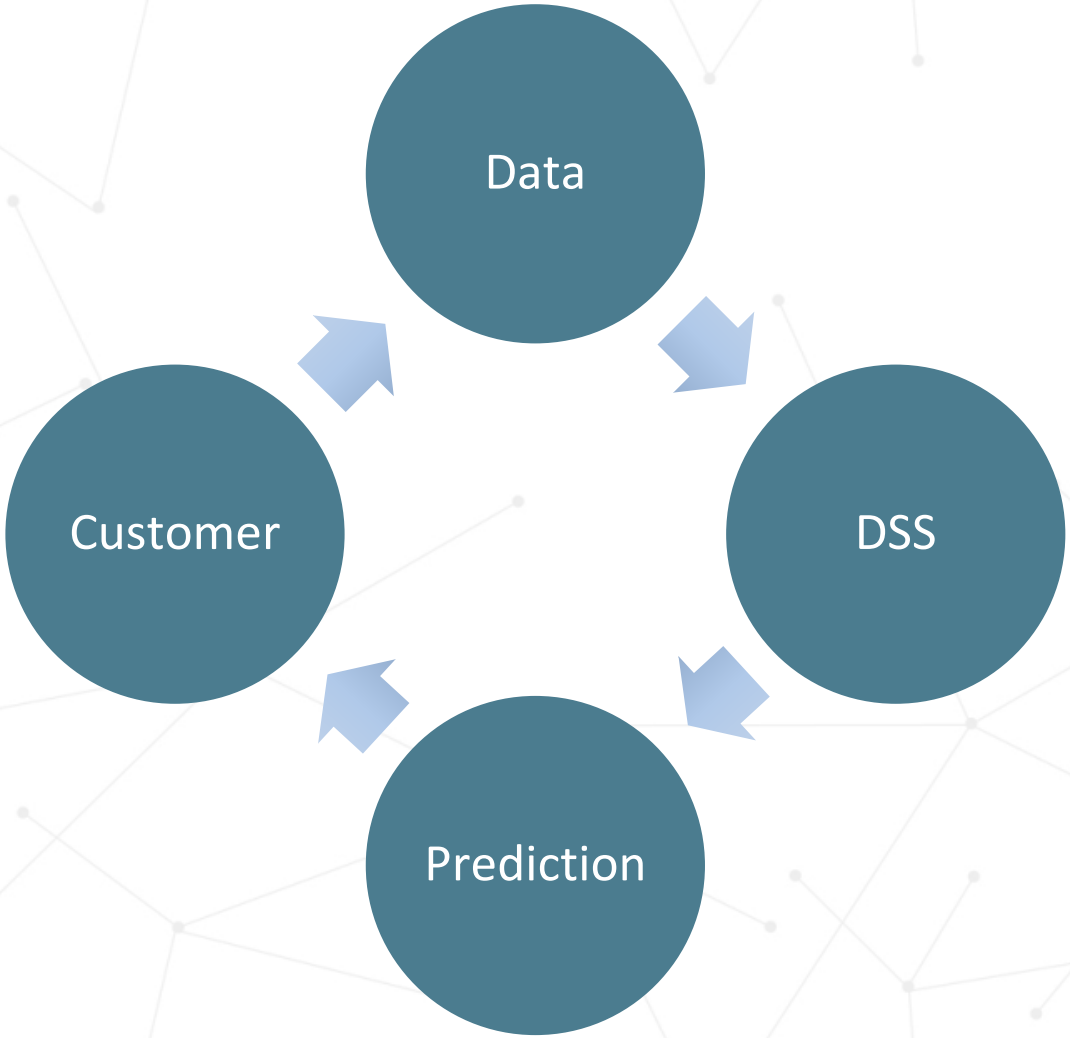
$ 25,500.00 per user, annually retail
$ 23,141.25 with 9.25% discount

Add

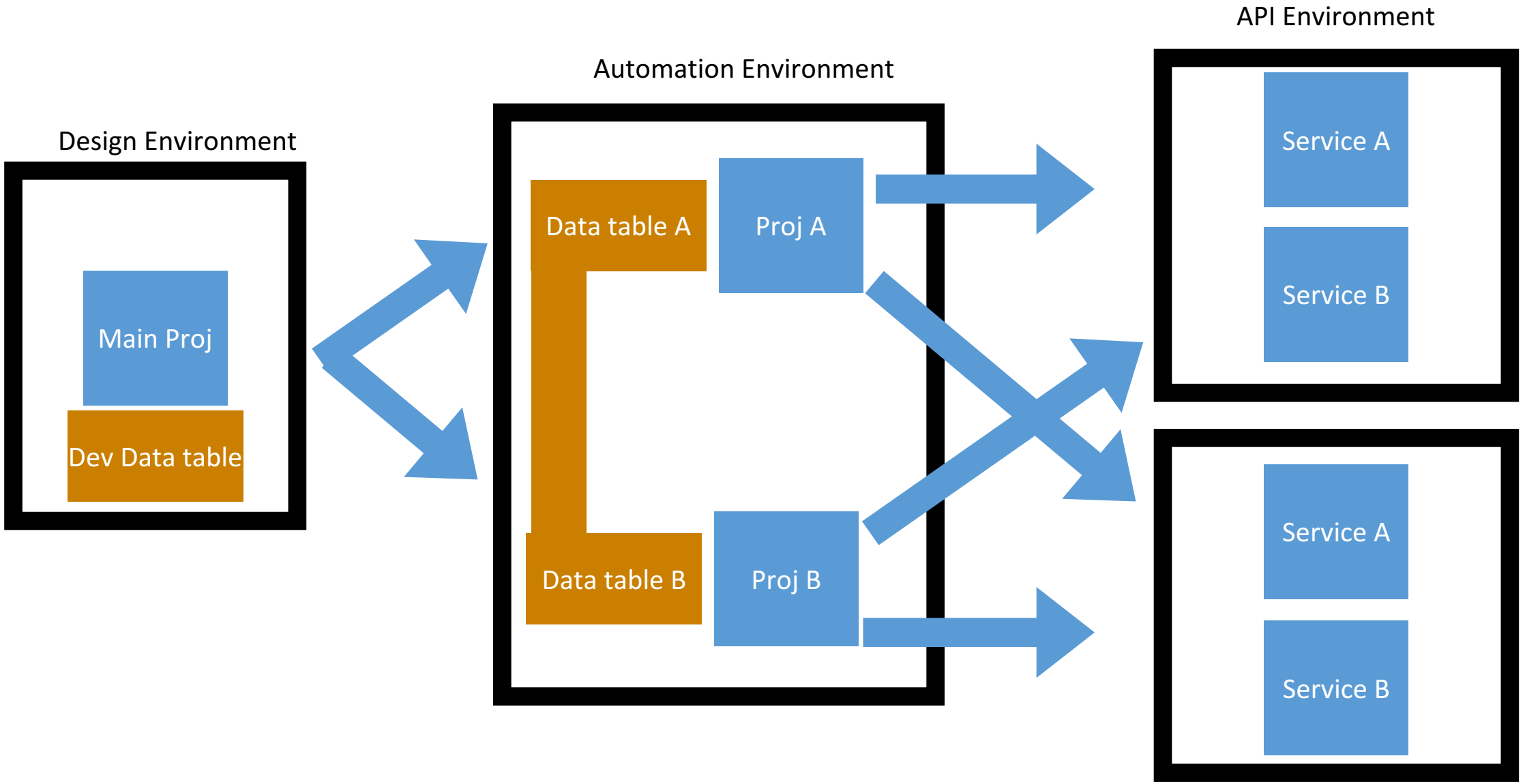
Recommendation

Layer 2 Switch is popular among customers with similar purchase history.

$ 1,225.00 each, retail
$ 1,086.57 with 11.3% discount


Data
DSS
Prediction
Customer

What had to be done?



One to Many Projects

* We must take one Project in Design and Turn it into many in Automation

Solution: Create a bundle and upload it to Automation multiple times

* We must connect to one ingestion dataset in Design and many in Automation

Solution: Connect to different databases on the Automation node

* We must turn one service in Design into many different services on the API nodes

Solution: Each new project generates its own service on the Automation node

A brief history of Dataiku and DSS

January 2013
Dataiku founded
Paris

February 2014
DSS 1.0
The 1st tool worldwide integrating visual data preparation and machine learning

January 2015
20 Employees
30 Customers
$4M Seed

April 2015
DSS 2.0
Real-time collaboration
Spark Integration

April 2015
Office in New York

October 2016
DSS 3.0
300% Yearly Growth
$14M Series A

February 2017
DSS 4.0
Office in London

August 2017
$28M Series B
100+ Employees
100+ Customers

Powering over 125 customers across 20 countries and multiple industries

Consumer Goods

Unilever L'ORÉAL

E-Retail

vente-privee traveloka

Consumer Electronics

hp gemalto

Travel

ACCORHOTELS.com trainline

Technology

KUKA GE

Transportation

DASSAULT FALCON AVIS

Financial Services

Santander AXA

Consulting

Capgemini McKinsey &Company

Healthcare

Pfizer

Media

npr UBISOFT

125+
Customers

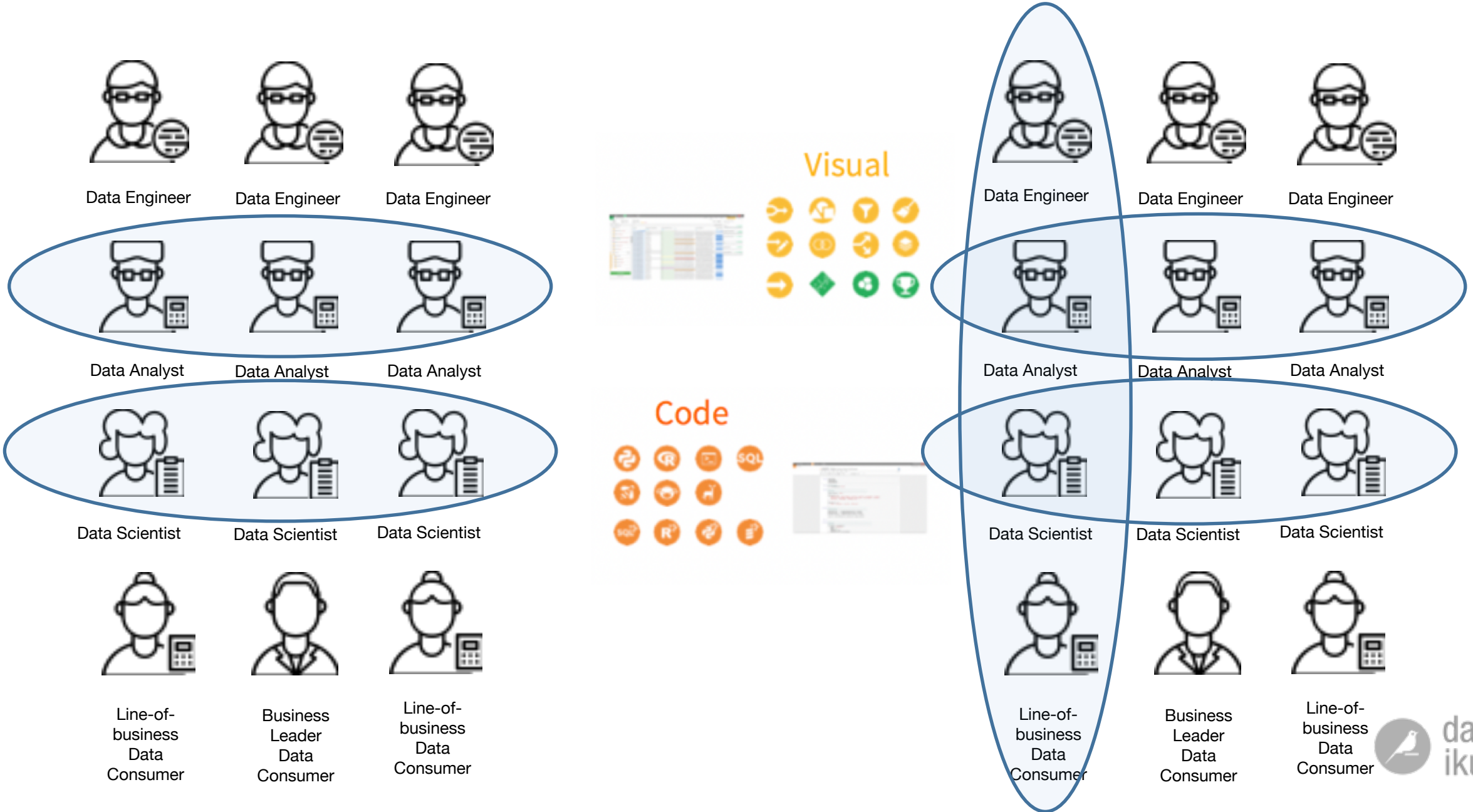
150%
Yearly Net Retention

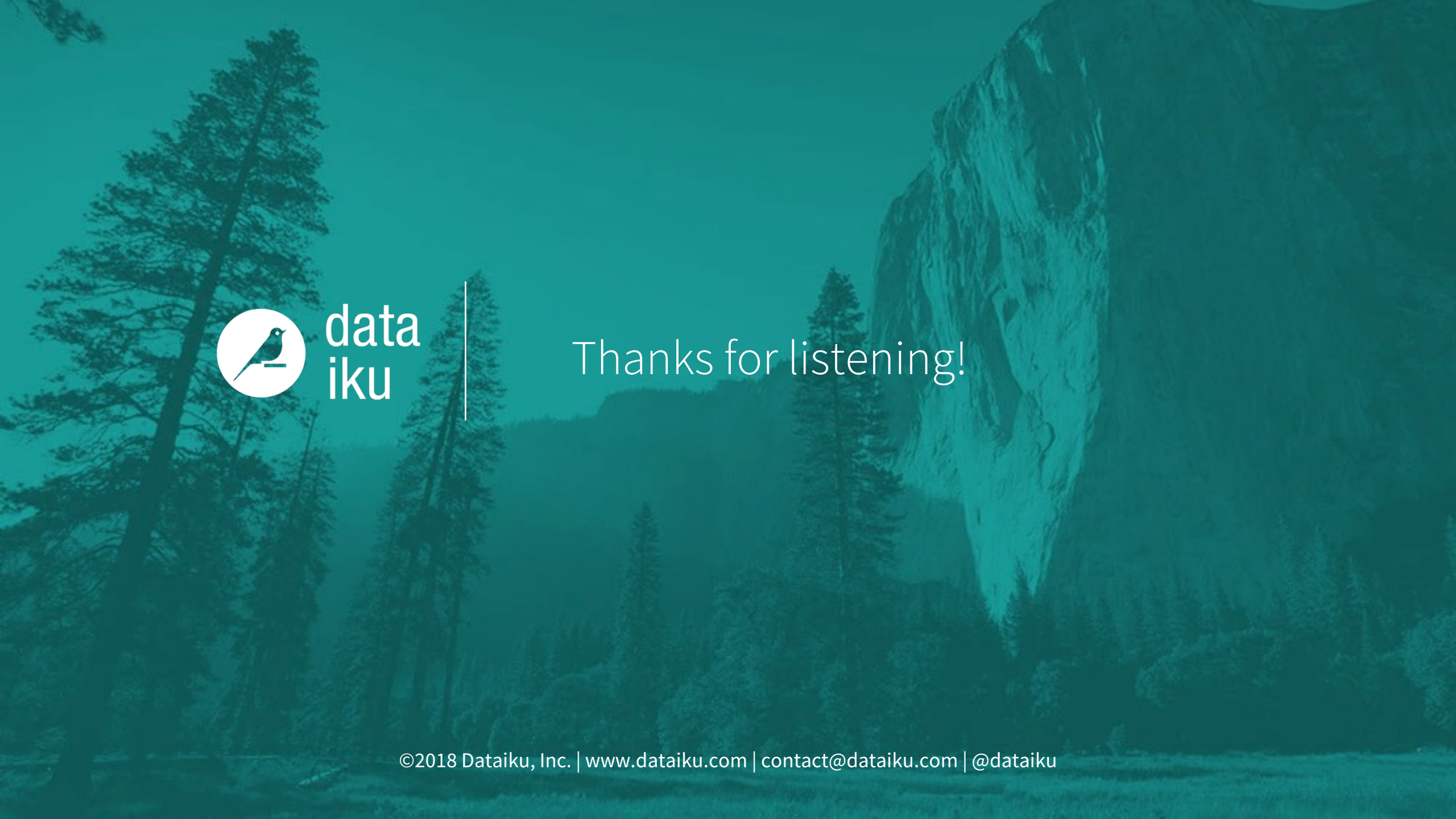
3X
YoY Growth
in 2017

Customers - Leaders in their industries

#1 Insurance Brand
#1 Pharma Brand
#1 Financial Information Company
#1 Flash Sales Company
#1 Car Sharing Company
#1 Cosmetics Company
#3 CPG Company

Horizontal Collaboration vs. Vertical Collaboration





Thanks for listening!