

Productionizing ml with FlowSpec

Subramaniam R, Danske Bank

#FlowSpec



Subramaniam Ramasubramanian

Sr. Data Scientist at Danske Bank. Working with BigData applications and Data since 2010

Copenhagen, Capital Region, Denmark



Danske Bank



Danmarks Tekniske
Universitet

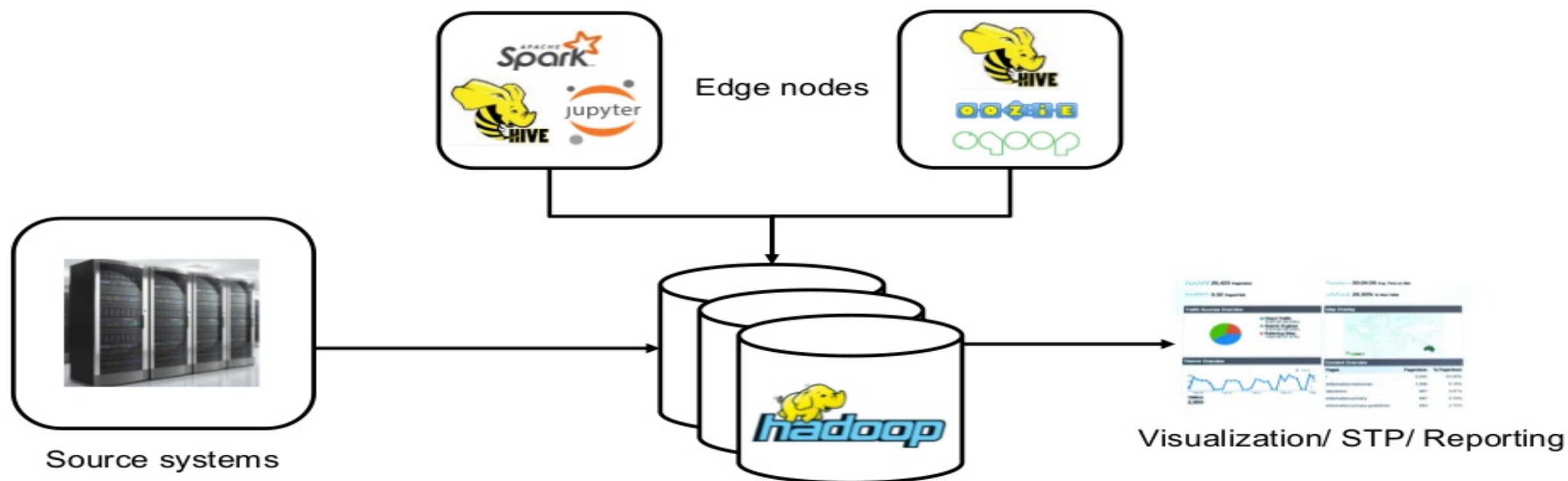
What can you take back from this presentation?

What can you take back from this presentation?




One possible way to get ML to production *fast!*

Some context

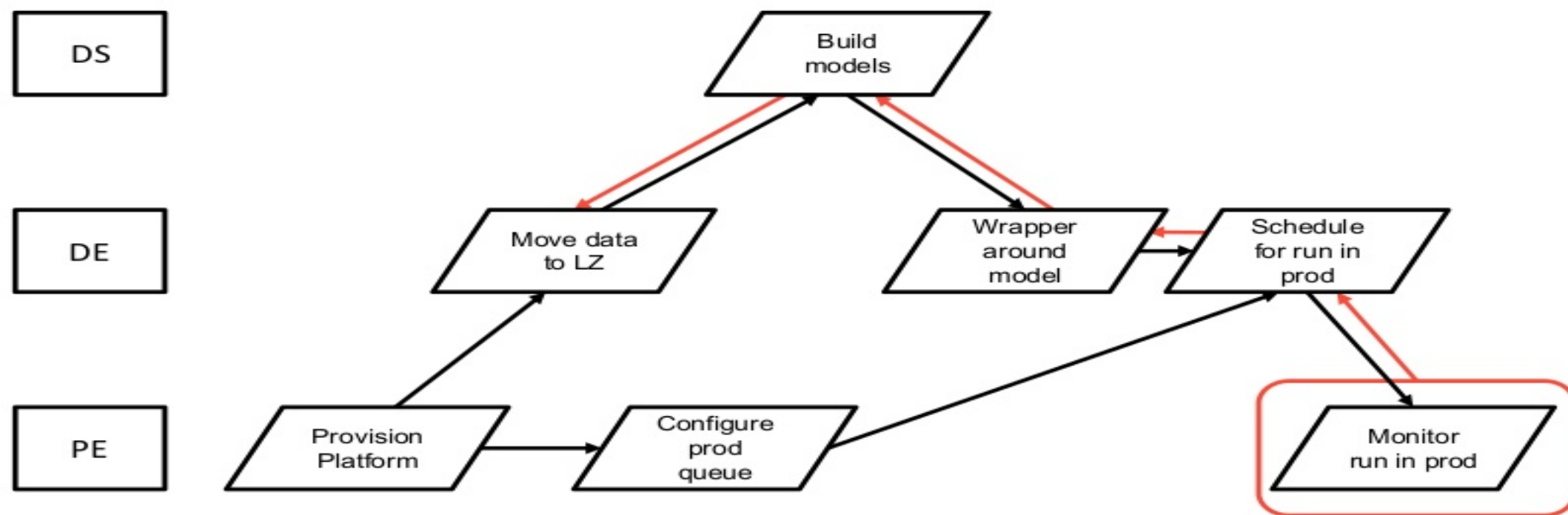
Infrastructure overview



Role description

Role	Tools	Questions they ask
DS		<ul style="list-style-type: none"> What is the scope for modeling/analytics? How do you build/evaluate these models? How/What features do you need? When am I going to actually do any of these?
DE		<ul style="list-style-type: none"> What are the reusable data products/transformations? Is the data moving to Hadoop? Is it validated? Wait, I am supposed to validate this?!
PE		<ul style="list-style-type: none"> What queue is this supposed to run on? Will upgrading break something? Are all the cluster services up? Why am I still in this job? Its 4AM.

Process flow

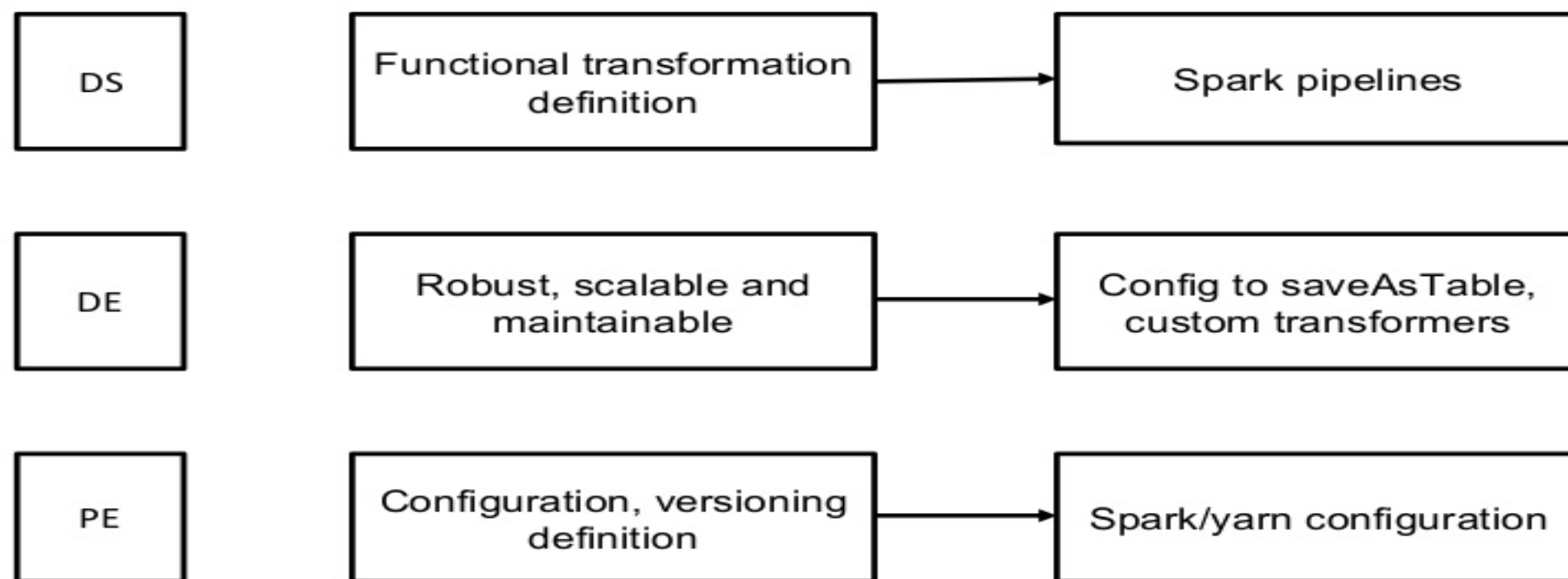


Everybody knows there's a
problem,

**Everybody knows there's a
problem, nobody knows who
should fix it!**

Enter FlowSpec

The ideal process



So what does that look like?

Demo

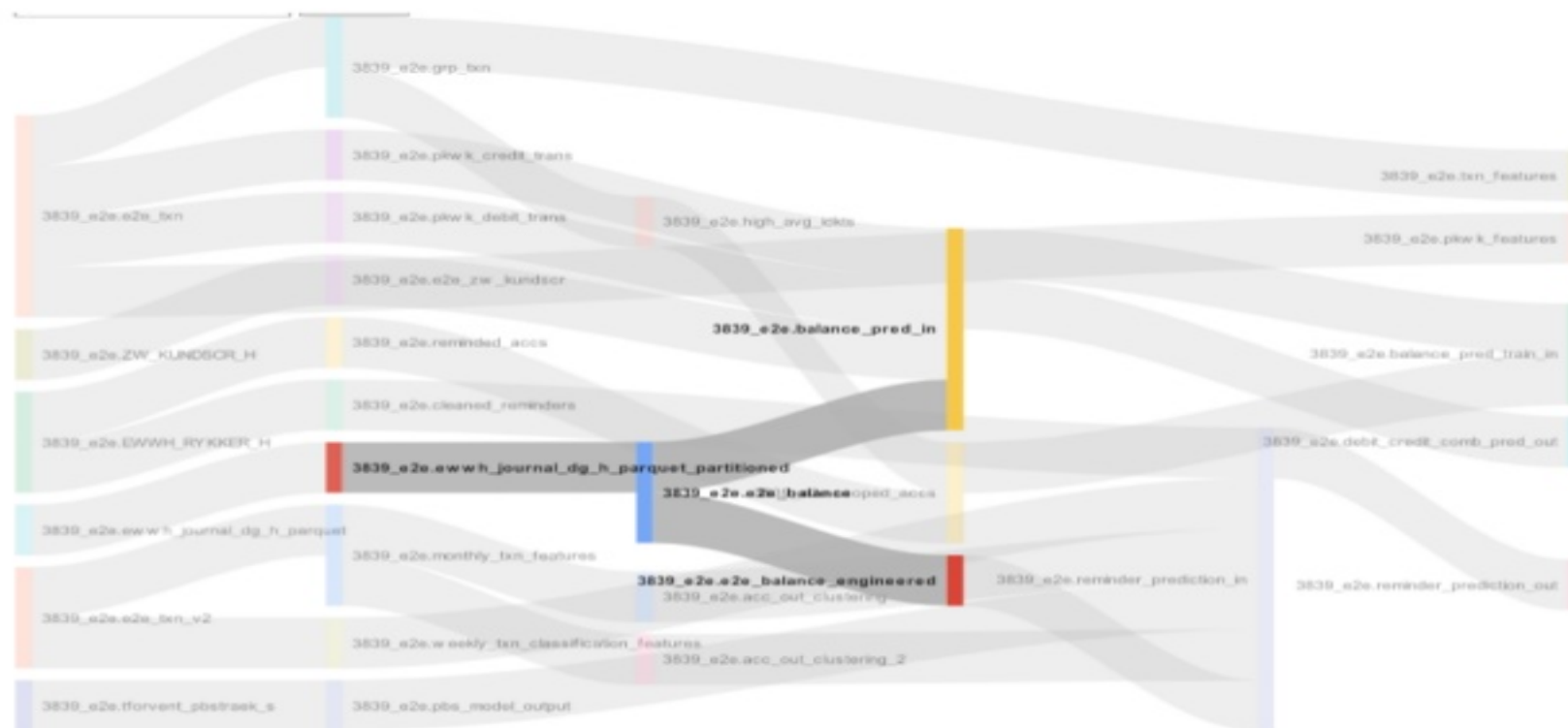
What it grows into

File name:	/home/BB4417/git/D6_end2end/moc	Table Created:	2018-07-16 21:09:19
Input Table:	3839_e2e_e2e_txn_v2	Table Updated:	2018-07-16 21:09:19
Output Table:	3839_e2e_weekly_txn_classification	Update Required:	True
Join Flag:	false	Input Exists:	True
Transformer File:	/projects/3839_e2e/feature_pipeline	Status:	None
Mandatory Columns:	knid, idkt, debit_flag, blps, bgdt_dt		
Join Conditions:			
Join Type:			
Column Prefix:			
Mode:	overwrite		
Partition spec:	dt_month		
Batch spec:			
Cluster Column:	idkt		
Cluster Spec:	200		
Cache Initial dataframe:	false		
Repartition Initial Dataframe:	-1		
File Format:	parquet		
Description:	The flow creates weekly account/customer level features based on txn classification for all personal banking DK customers.		

[Explore](#)
[Update Table](#)

[Save](#)
[Update with Dependencies](#)

What it grows into



Future work

- Include more data validation types
- Improve command line interface
- Create a culture for custom transformers
- Push for more deliveries through pipelines
- Possibly make FlowSpec open source

Summary

- Use Spark pipelines to abstract functionality
- Reuse code through transformers
- Simplify distributed computing for DS
- Streamline the process with FlowSpec

Thank you!