Introducing Arc

Engineers Shouldn’t Write ETL

Engineers Shouldn't Write ETL: A Guide to Building a High Functioning Data Science Department

JEFF MAGNUSSON

March 16, 2016 - San Francisco, CA

Tweet this post! Post on LinkedIn

"What is the relationship like between your team and the data scientists?" This is, without a doubt, the question I'm most frequently asked when conducting interviews for data platform engineers. It's a fine question – one that, given the state of engineering jobs in the data space, is essential to ask as part of doing due diligence in

The Four ETL Roles

Users aka ‘the Money’	Data Scientists aka ‘the Thinkers’	Data Engineers aka ‘the Doers’	Infrastructure aka ‘the Plumbers’
Understand the customers and business so can define requirements.	Better engineers than statisticians and better statisticians than engineers.	Build pipelines that feed the users and data scientists with data and take the ideas from the users and data scientists and implement them.	Maintain the databases / clusters / big data infrastructure.

The Four ETL Roles: Problems

Users aka ‘the Money’	Data Scientists aka ‘the Thinkers’	Data Engineers aka ‘the Doers’	Infrastructure aka ‘the Plumbers’
Why does it take so long to implement?	Pandas is the best thing ever.	This is super boring, I am a Thinker.	I can’t reproduce the Python environment.
Why is this number different in this other report?	My personal preference is this Machine Learning library.	This machine learning model doesn’t even run.	How do I upgrade do I have to rebuild JARs?
	It works on my laptop.	This R Model using native C++ library doesn’t work on my cluster.	This doesn’t scale.

Introducing Arc