Use of standards and related issues in predictive analytics

KDD 2016, SF 2016-08-16

Paco Nathan, @pacoid
Dir, Learning Group @ O'Reilly Media

PMML referenced by 86 publications in Safari, 2001-2016

https://www.safaribooksonline.com/search/?query=PMML

Search for : Safari

https://www.safaribooksonline.com/search/?query=PMML&sort=publication_date&highlight=true&is_academic_institution_account=false&extended_publisher_d...

O'REILLY®

Programming Hive
by Edward Capriolo, Dean Wampler, Jason Rutherglen
Publisher: O'Reilly Media, Inc.
Published: September 2012
6 Reviews

23. Case Studies (73 mins)

If you are interested in seeing what others think and would like to participate, please visit: http://karmasphere.com/hive-features-survey.html [26] http://www.r-bloggers.com/taking-r-to-the-limit-large-datasets-predictive-modeling-with-**pmml**-and-adapa/ [27] For more details on the use of generalized additive models (GAM), see Hastie et al. 2001.

Advancing Technologies and Intelligence in Healthcare and Clinical Environments Breakthroughs
by Joseph Tan
Publisher: IGI Global
Published: June 2012
0 Reviews

CONCLUSION (7 mins)

Predictive Model Markup Language (**PMML**) version 3.0 specification. The approach taken in this research represents the mined knowledge as **PMML** models which are self-describing XML documents. ... Retrieved August 21, 2007, from http://www. dmg.org/**pmml**-v3-0.html Duch, W., Gr , abczewski, K., Adamczak, R., & Grudzi'nski, K. (2001).

E-Marketing
by Information Resources Management Association, USA
Publisher: IGI Global
Published: May 2012
0 Reviews

Chapter 17: Database Marketing Process Supported by Ontologies (50 mins)

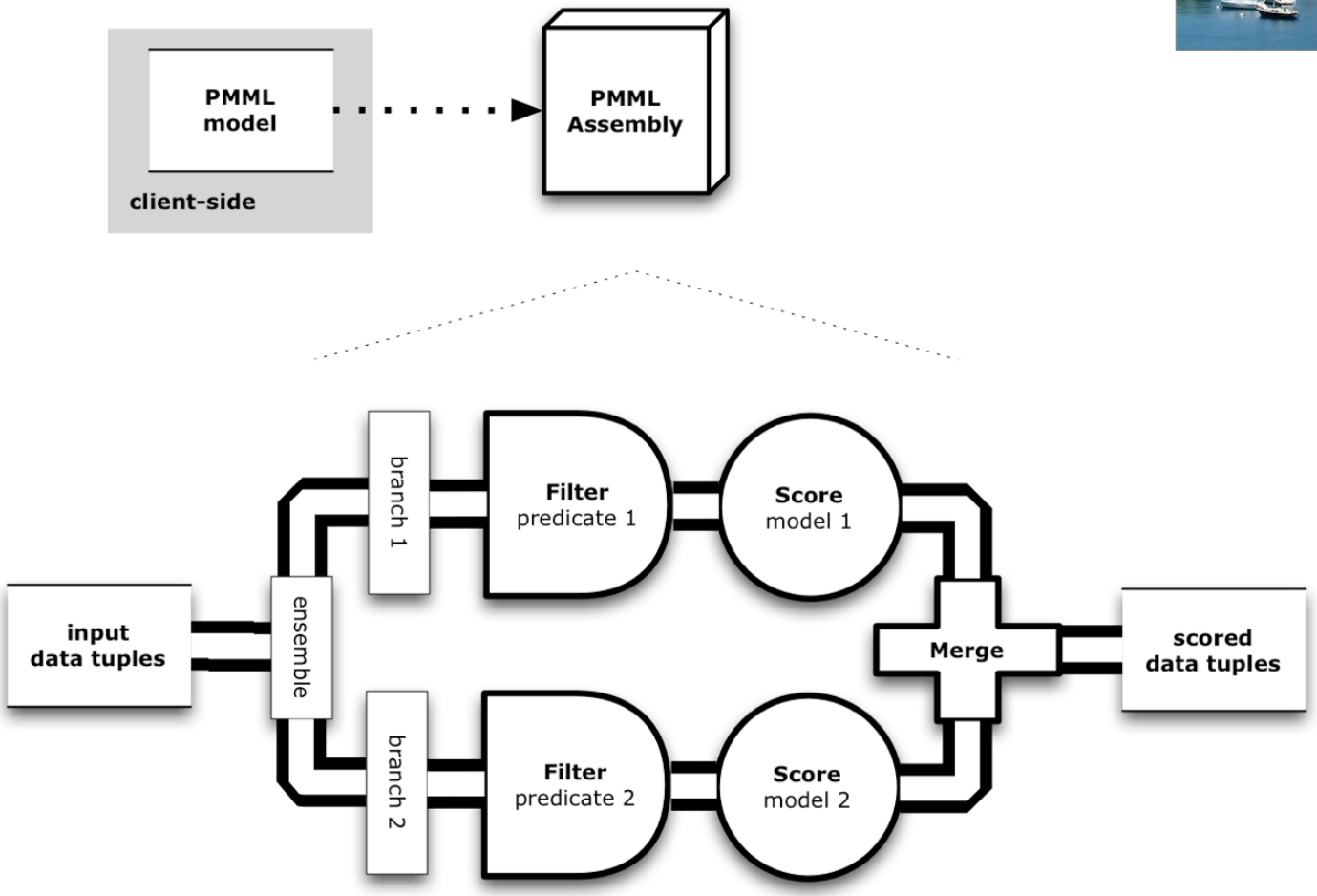
PREDICTIVE MODEL MARKUP LANGUAGE Predictive model markup language (**PMML**) is an XML-based language that provides a way for applications to define statistical and data mining models and to share these models between **PMML** compliant applications (Data Mining Group). ... Such DM-elements are represented by OWL classes together with variations of their representations in XML (allowing information interchange with **PMML** DM models). Since **PMML** is an XML based standard, its specification comes in the form of an XML schema that defines language primitives as follows (Brezany et al., 2008): Data Dictionary: It defines fields that are the inputs for models and specifies their types and value ranges.

Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics
by Bill Franks
0 Reviews

Pattern: PMML for Cascading and Hadoop

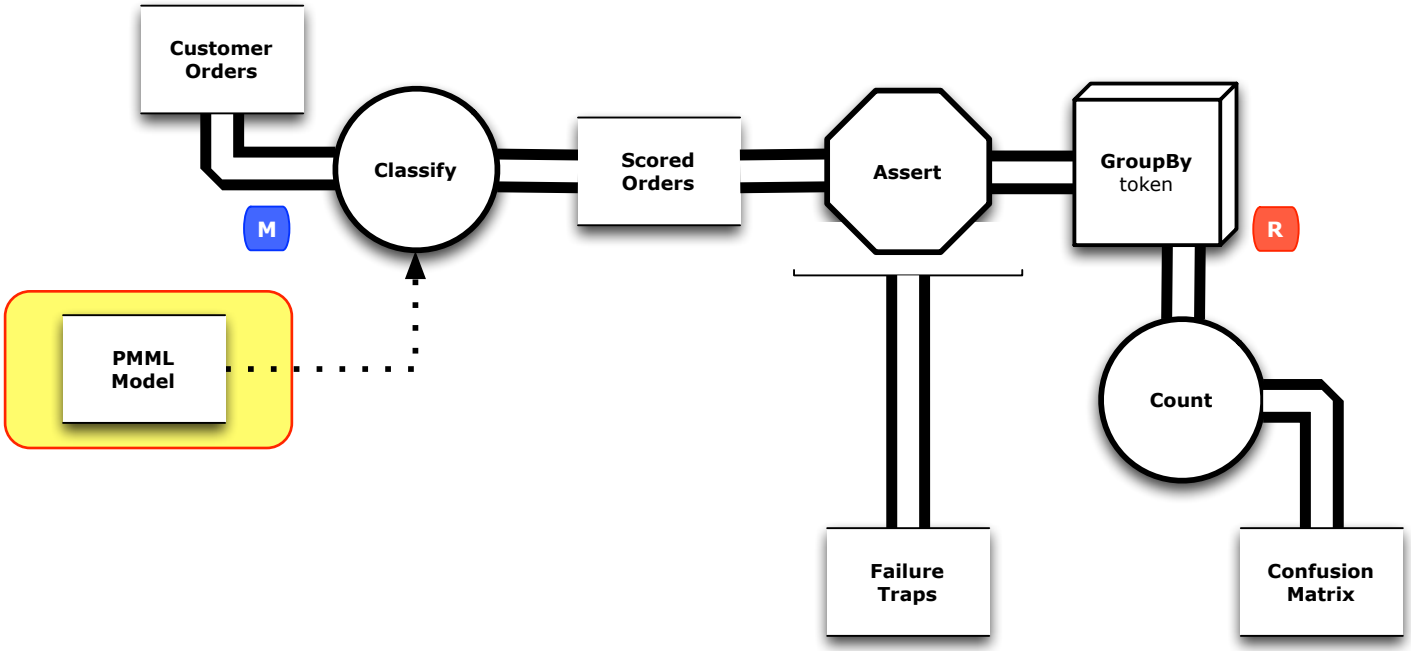
P Nathan, G Kathalagiri (2013-08-11)

https://goo.gl/jk7829



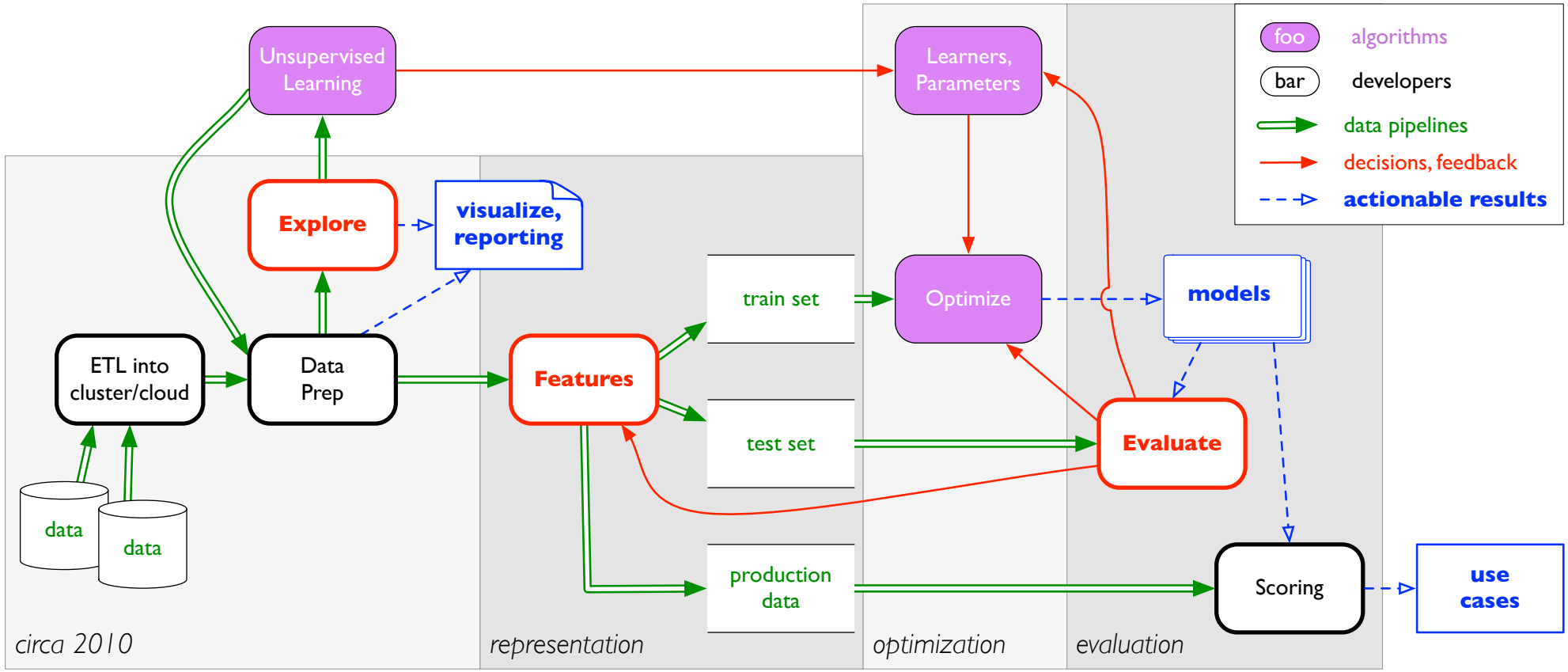
Pattern – score a model, using pre-defined Cascading app

cascading.org/projects/pattern



Generalized Workflow for ML Use Cases in Big Data

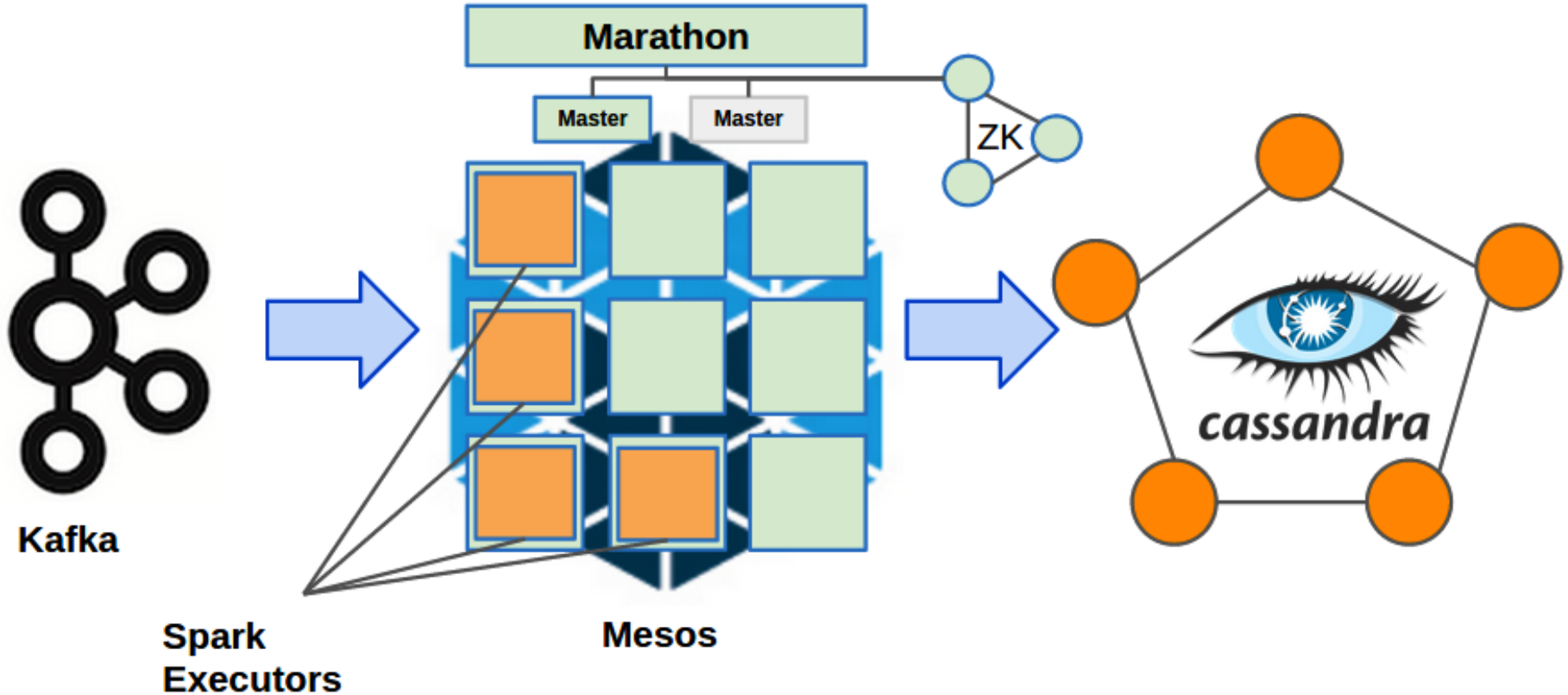
Algorithms and developer-centric template thinking only go so far in real-world workflows…



Results shown in blue, hard problems highlighted in red

“One Size Fits All” Doesn’t Anymore

This common architectural pattern requires interchange…



Tuning Spark Streaming for Throughput
Gerard Maas, Virdata (2014-12-22)

IoT alters “velocity” and “volume” dramatically

This growing category of use cases requires interchange…

bits.blogs.nytimes.com/2013/06/19/g-e-makes-the-machine-and-then-uses-sensors-to-listen-to-it/

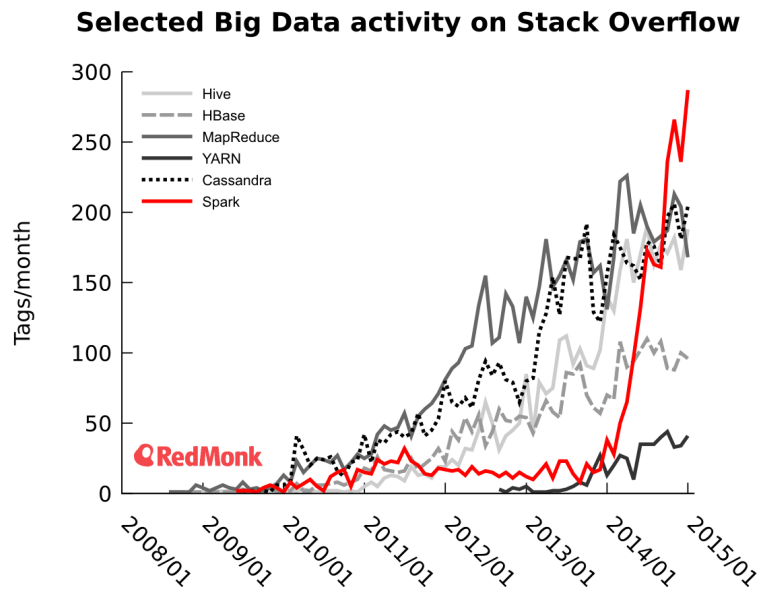
Lessons from the success of Apache Spark…

interchange is necessary for the ecosystem

major use cases tend to build their own ML libraries – despite a case where a majority of committers tend to support a common vision and encourage use of a canonical library (MLLib with DataFrames)

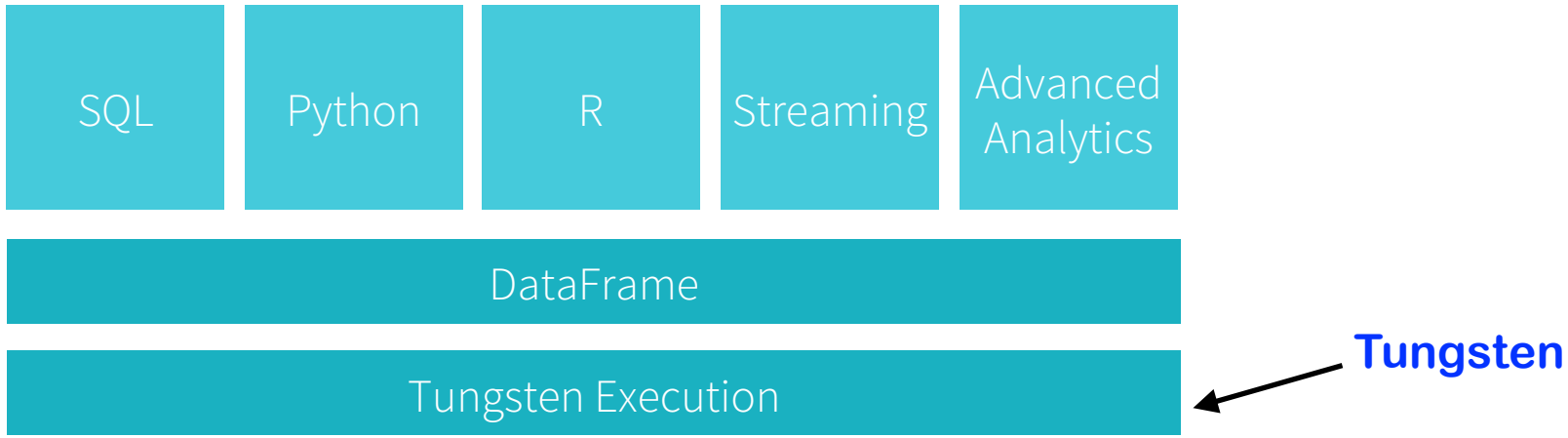
when a successful business grows over time, challenges arise by definition: managing separated teams, mergers and acquisitions, increased audits, regulations, etc.

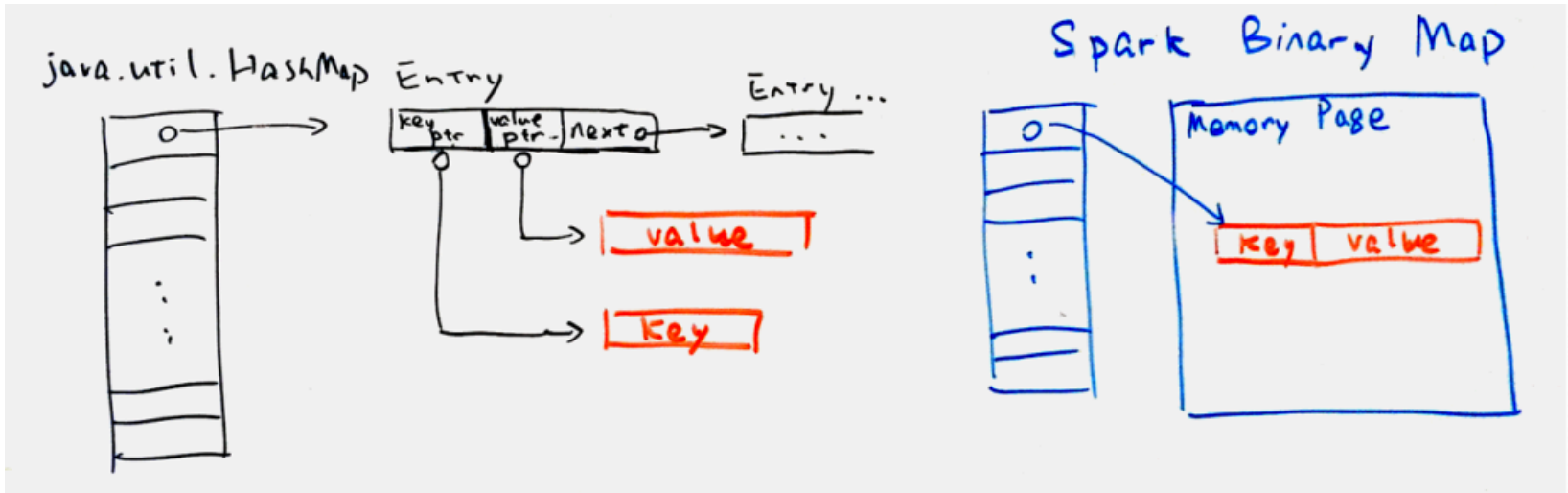
therefore, lack of interchange for analytics represents a serious technical debt and potential liability



Lessons from the success of Apache Spark…

direct use of "compilers" becomes atypical as abstraction layers become smarter for deferred optimization





What to suggest for existing standards?

microservices: how to compose models + parameters from multiple/distinct services

support for API definitions in Swaggar http://swagger.io/

consider the benefits of Parquet, e.g., how pushdown predicates enable better optimization of workflows

What to suggest for existing standards?

additional standards emerging for other aspects of workflow definition:

Jupyter http://jupyter.org/

create and share documents that contain live code, equations, visualizations and explanatory text

Dat http://dat-data.com/

shares versioned data through a decentralized network

What to suggest for existing standards?

other lingering issues:

data lineage / provenance

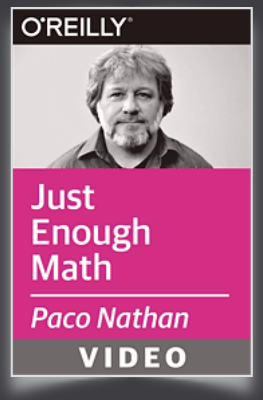
metadata drift

https://public.resource.org/about/

presenter:

monthly newsletter for updates, events, conf summaries, etc.:

liber118.com/pxn/



Just Enough Math
O'Reilly (2014)
justenoughmath.com