Configurable Data Pipelines

Towards a Reference Architecture

E. U. Syed
Philips Lighting Research
Sunday, June 18, 2017

innovation + you

JADS Jheronimus Academy of Data Science

PHILIPS

A Little About Me

Ekhtiar Syed

Scientist/ Data Engineer,
Philips Lighting Research

Graduated International Masters in Service Engineering
With Cum Laude in July 2016

Universität Stuttgart

UNIVERSITY OF CRETE

TILBURG UNIVERSITY

- Distributed Systems
- Big Data Solutions
- Internet of Things
- Machine Learning & Data Mining

JADS PHILIPS

Data and Crude Oil

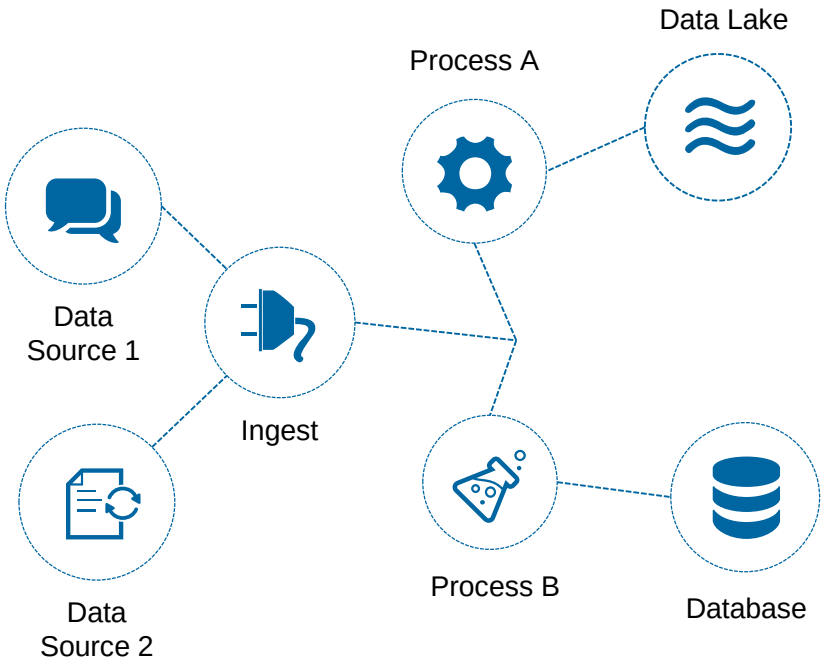
In 2006, Michael Palmer, a marketing commentator was the first to compare Data with crude oil.

Just like crude oil, data may turn into a valuable commodity through proper processing.

Data pipelines constitute the refineries of data: they efficiently collect, process, and transform raw data into actual value.

P. Rotella, "Is Data The New Oil," Forbes, 02 April 2012. [Online]. Available: http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/. [Accessed 04 April 2016]

What is a Data Pipeline?

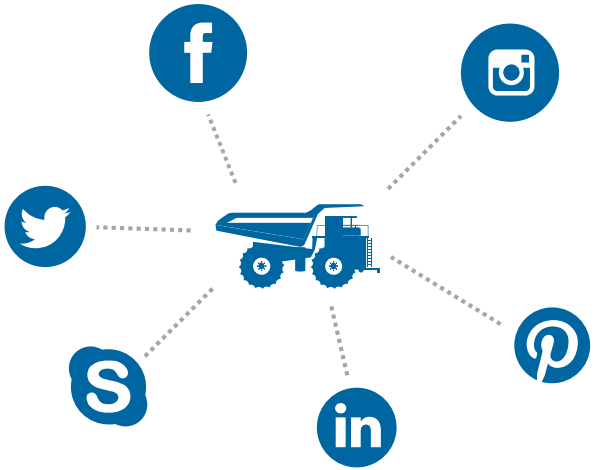


Traditionally, a pipeline is a collection of data processing tasks connected in a series, where the output of one task is the input of the next task. [1]

Data pipelines in real-world settings typically consist of multiple tasks leveraging different technologies to meet required design goals or considerations.

[1] in Encyclopedia Of Information Technology, New-Delhi, Atlantic Publishers & Dist, 2007 , p. 382.

Data Pipeline in Companies



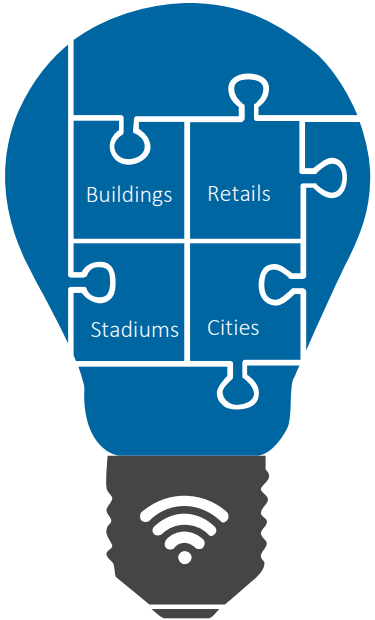
Netflix has a **data pipeline** to process **1.3 petabyte** of data per day to enable features like movie recommendation [1].

Facebook's real time data pipeline powers use cases like insights for Facebook page and analytics for mobile applications [2].

Twitter has a data pipeline to use **deep learning** at scale and show the **best Tweets** for your timeline [3].

[1] https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905 [2] https://research.fb.com/publications/realtime-data-processing-at-facebook/ [3] https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html

Data Pipeline and Philips Lighting



#1 Manufacturer of lighting products with history of 125 years

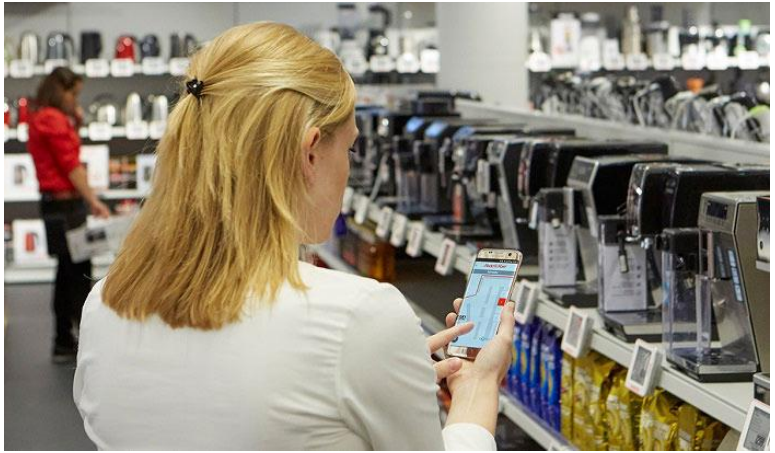
#1 in Connected Lighting with further intention of innovation

Obvious Data Enabled Use Cases
Easy Maintenance, Energy Savings

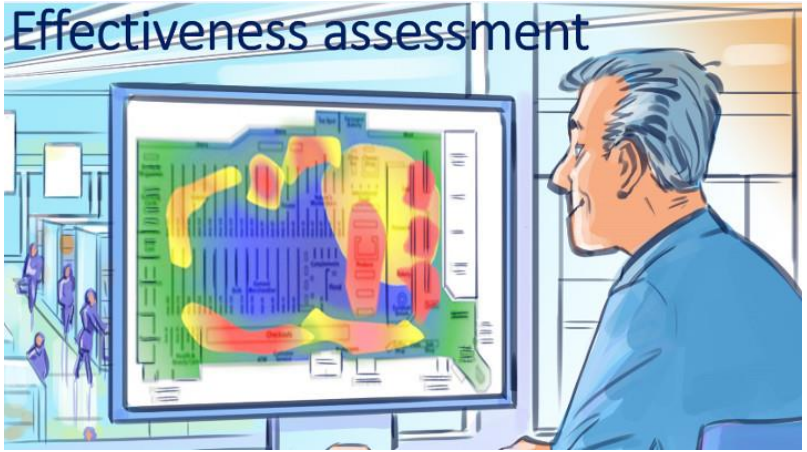
Hidden Data Enabled Use Cases
Shopper Analytics

Source: http://www.lighting.philips.com/main/company/beyond-illumination/internet-of-things

Data Pipeline and Philips Lighting



Product Finding



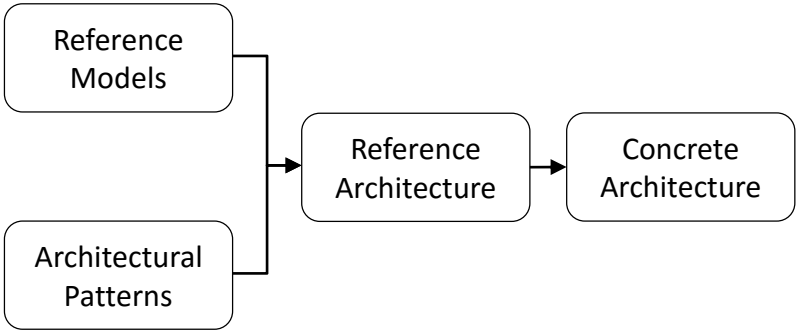
Analyze Shopper Traffic & Behavior

http://www.lighting.philips.com/main/systems/themes/led-based-indoor-positioning

Research Question

What are the **components** and **data flow** for a **reference architecture** of a (big) **data pipeline**.

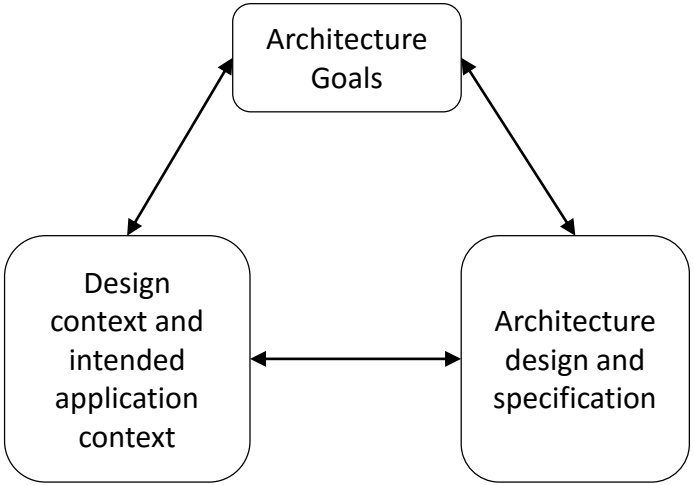
Definition - Reference Architecture



- Defines the division of functionalities and data flow in between the components.
- provides a high level abstraction to facilitate concrete architectures
- Derived from architectural patterns and reference models

P. G. D. G. Samuil Angelov, "A framework for analysis and design of software reference architectures," *Information and Software Technology,* vol. 54, no. 4, pp. Pages 417-431, April 2012.

A Framework for Analysis of Reference Architecture



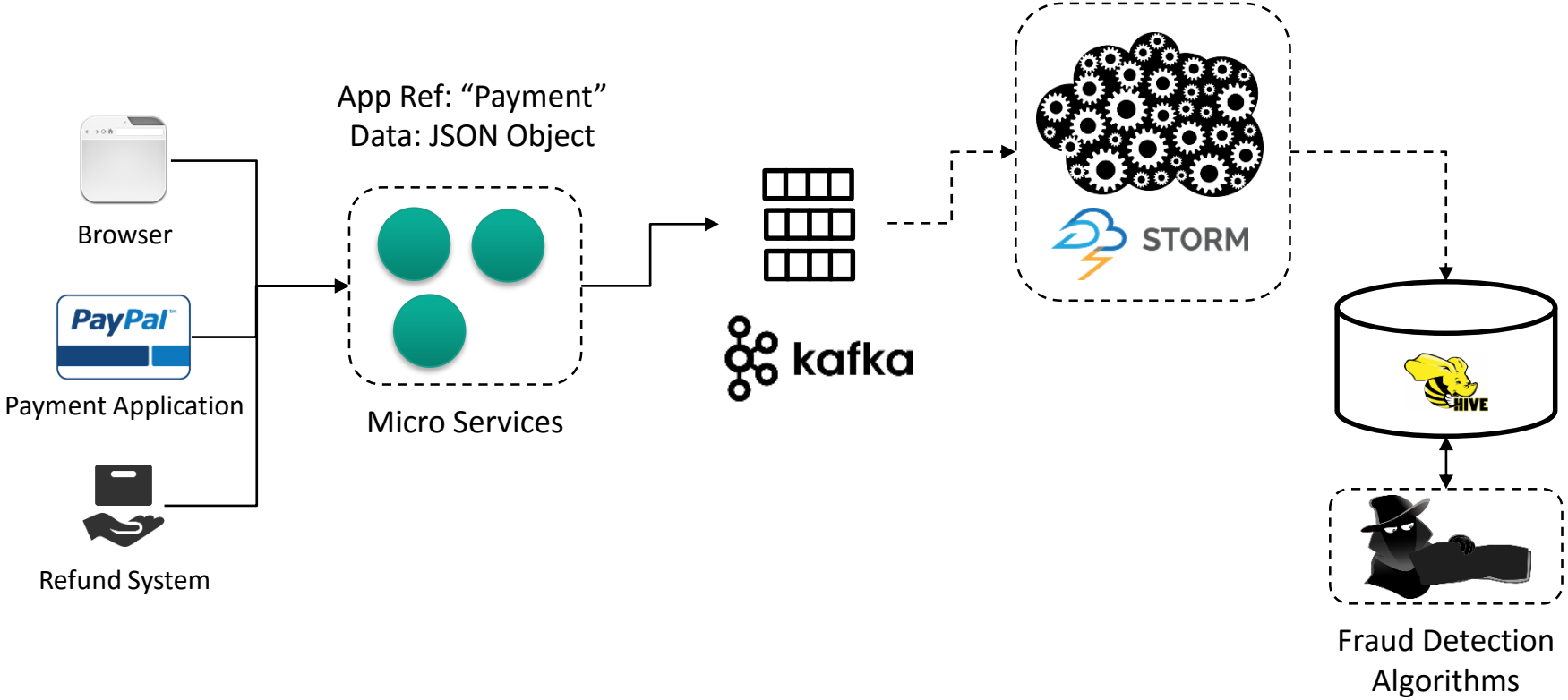
	Sub Dimensions	Values
G1: Why	Why is it defined?	Standardization / facilitation
C1: Where	Where will it be used?	Single / multiple organizations
C2: Who	Who defines it?	Research cent., software org., user org., standardization org..
C3: When	When is it defined?	Preliminary / Classical
D1: What	What is described?	Components and connectors, interfaces, protocols, algorithms, policies..
D2: How	How detailed is it described?	Detailed/ semi-detailed/ aggregated.
D3: How	How concrete is it described?	Abstract/ semi-concrete/ concrete
D3: How	How is it represented?	Informal/ semi-formal/ formal

P. G. D. G. Samuil Angelov, "A framework for analysis and design of software reference architectures," *Information and Software Technology,* vol. 54, no. 4, pp. Pages 417-431, April 2012.

Reference Architecture Design Approach

- Several data pipeline use cases from industries are reviewed.
- A data pipeline is developed and deployed part of an experiment.
- Architectural Patterns (or anti patterns) are extracted.
- These patterns are used to formulate the reference architecture
- Analyze congruency of reference architecture in multidimensional space

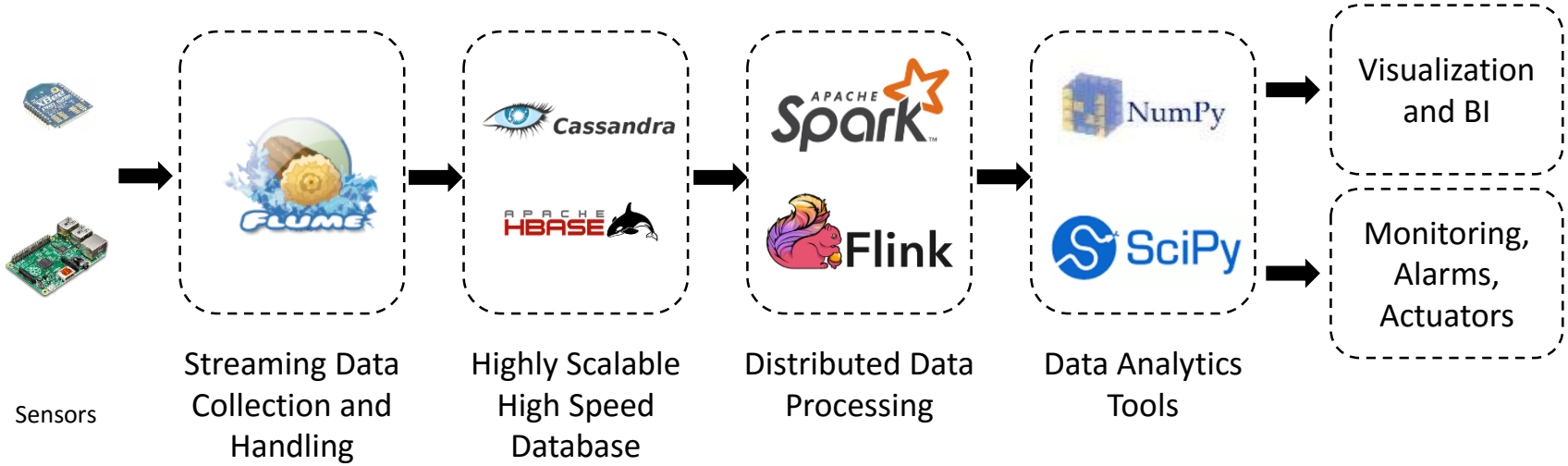
PayPal: Carrier Payments Data Pipeline



M. Mahendru, "PayPal Engineering," PayPal , 2016 November 15. [Online]. Available: https://www.paypal-engineering.com/2016/11/15/carrier-payments-big-data-pipeline-using-apache-storm/. [Accessed 10 March 2017].

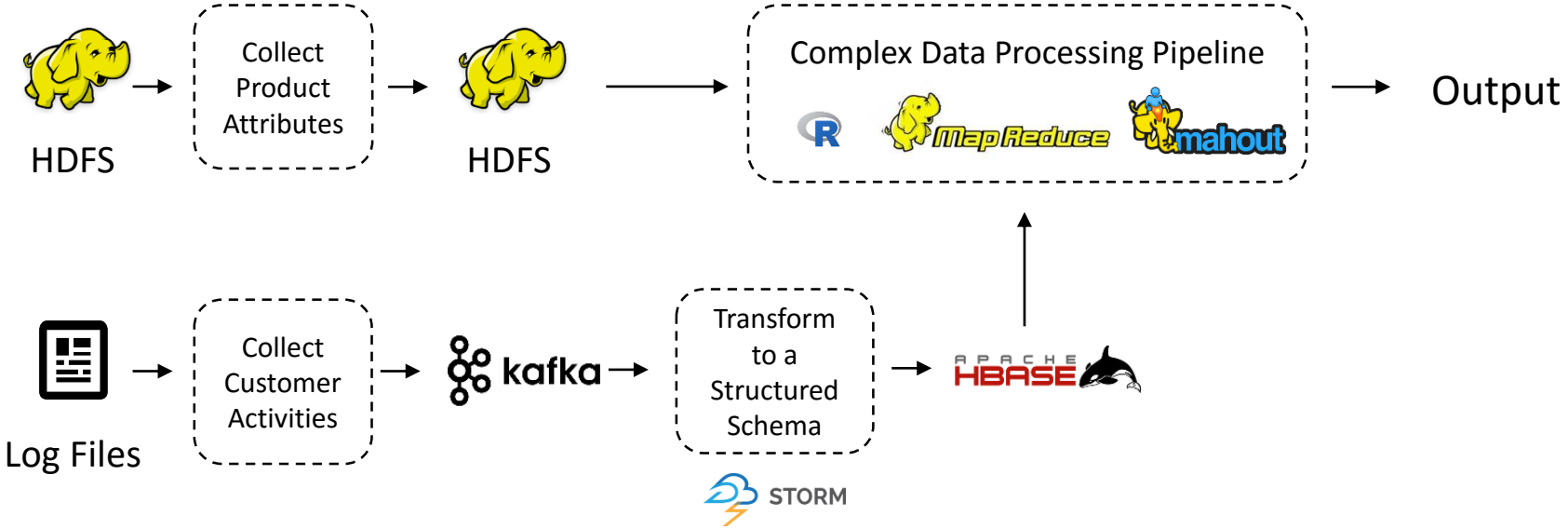
JADS PHILIPS

Data Pipeline for Pervasive Sensor



A. I. Jussi Ronkainen, "Designing a data management pipeline for pervasive sensor communication systems," in The 12th International Conference on Mobile Systems and Pervasive Computing, Belfort, 2015.

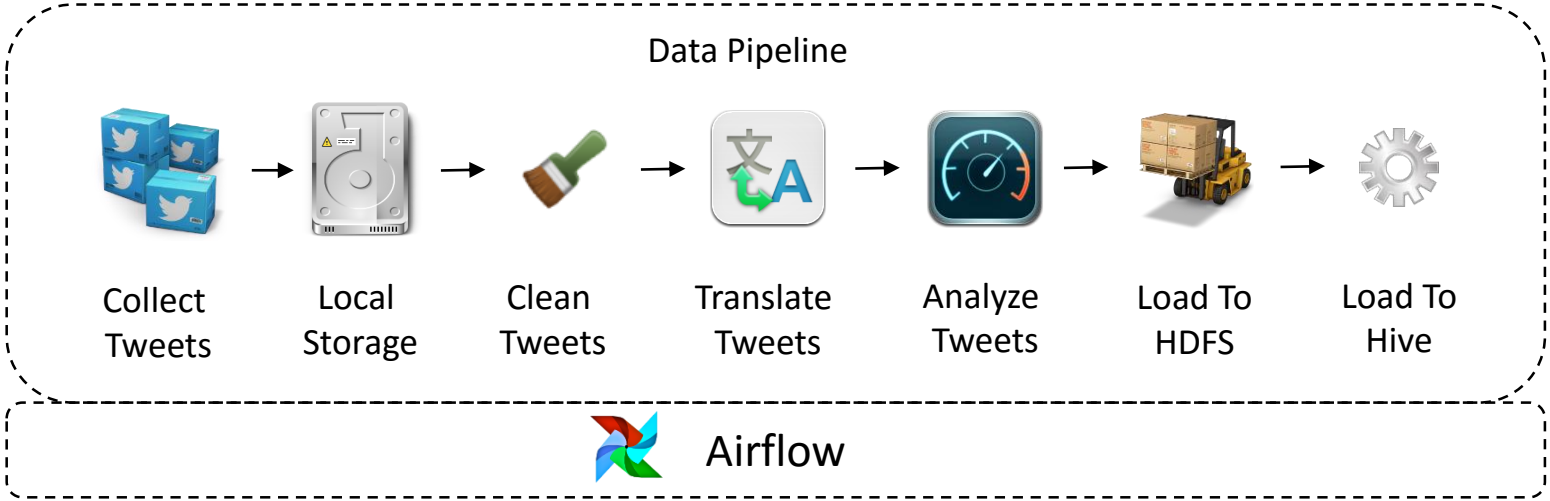
Groupon: CRM Data Gathering and Mining Pipelines



V. D. a. N. P. Kang Li, "Big Data Gathering and Mining Pipeline for CRM using Open-source," in *IEEE International Conference on Big Data*, Dalian, 2015.

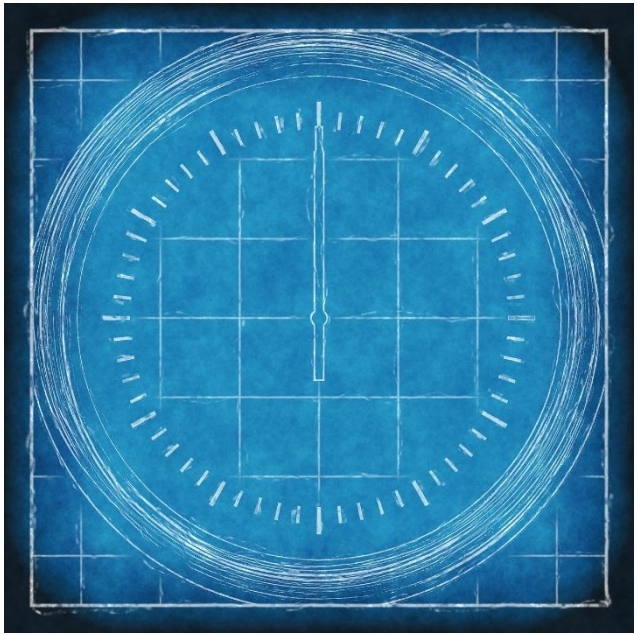
JADS PHILIPS

Analyzing Customer Sentiment



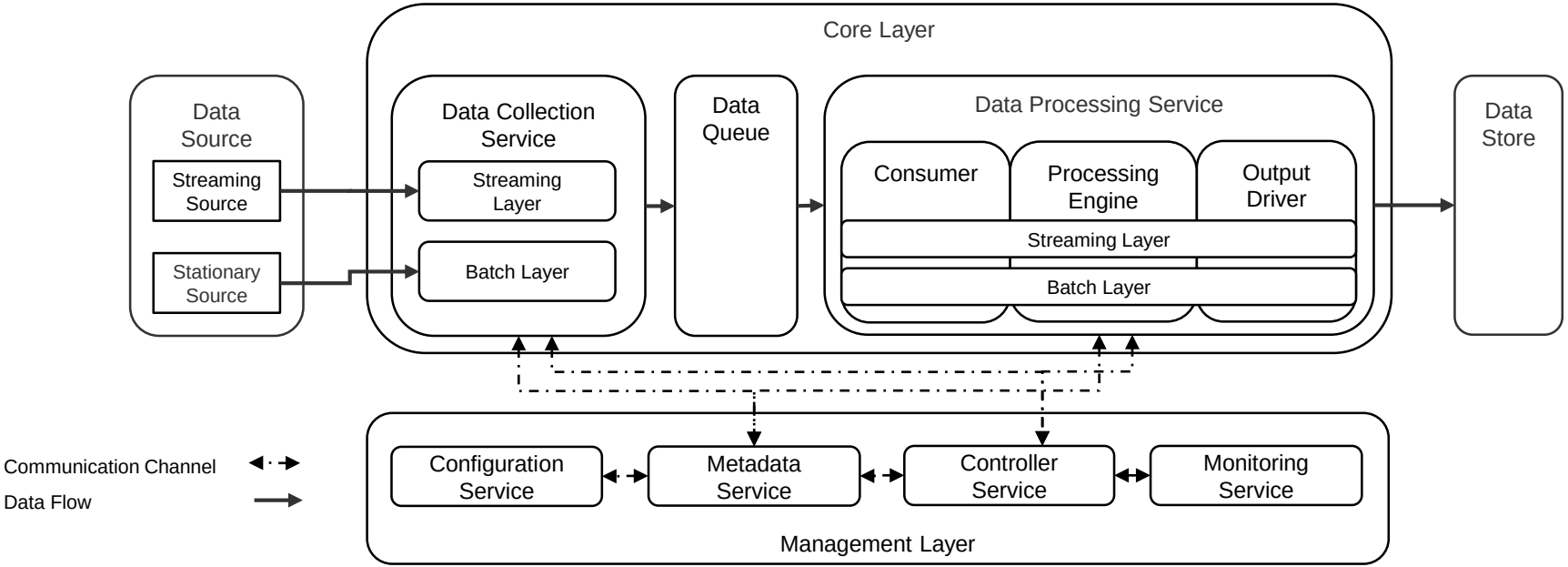
E. U. Syed, "DamStream: A Real Time Data Pipeline Framework For Realizing Automated Data Lakes," Tilburg University Library, Tilburg, 2016.

Summarizing Architectural Patterns



	PayPal	Groupon	IoT App.	Twitter Sentiment
Schema Flexibility	●	●		
Real-Time Capabilities	●		●	
Complex Processing		●	●	
System Scalability	●	●	●	●
Pub/Sub Mechanism	●		●	
Microservices	●			
Controller				●
Single Point Configuration				●

Reference Architecture



Further Analysis of Reference Architecture

Type :5 (Variant 5.1) [1]

A preliminary, facilitating reference architecture is designed for multiple organizations by a research center.

Futuristic design, doesn't concentrate on the requirements but on the innovative elements of the architecture.

Their main contribution is in inspiring future research efforts in the domain.

	Sub Dimensions	Values for our RA
G1: Why	Why is it defined?	Facilitation
C1: Where	Where will it be used?	Multiple Organizations
C2: Who	Who defines it?	Research Center
C3: When	When is it defined?	Preliminary
D1: What	What is described?	Components, Data Flow
D2: How	How detailed is it described?	Semi-Detailed
D3: How	How concrete is it described?	Abstract elements
D3: How	How is it represented?	Semi-formal element specifications

Type :5 (Variant 5.1) [1]

[1] P. G. D. G. Samuil Angelov, "A framework for analysis and design of software reference architectures," *Information and Software Technology,* vol. 54, no. 4, pp. Pages 417-431, April 2012.

JADS PHILIPS

Conclusion & Future Work

- Only Limited to three case studies and one internal experiment
- Increase the number of use cases in future with recent publications
- The reference architecture was only analyzed theoretically
- Practical implementation involving multiple organizations

Questions and Feedbacks

PHILIPS

PHILIPS