

Performance study of Spindle, a web analytics query engine implemented in Spark CloudCom 2014

Brandon Amos* and David Tompkins
Adobe Research

*Adobe intern, Ph.D. Student at Carnegie Mellon University.

December 19, 2014

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking concurrent queries.

Scaling Spark and HDFS workers.

Future Work

Conclusions

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking concurrent queries.

Scaling Spark and HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

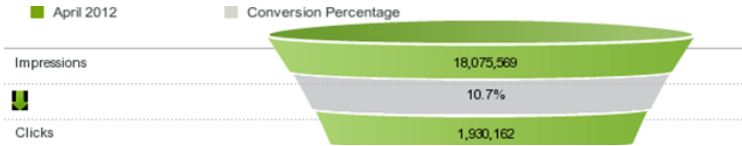
Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

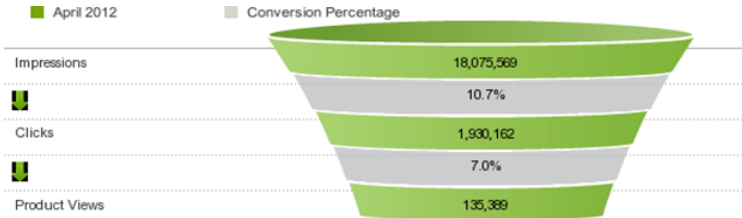
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

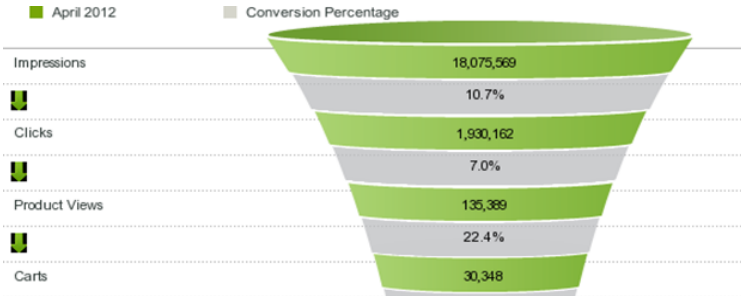
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

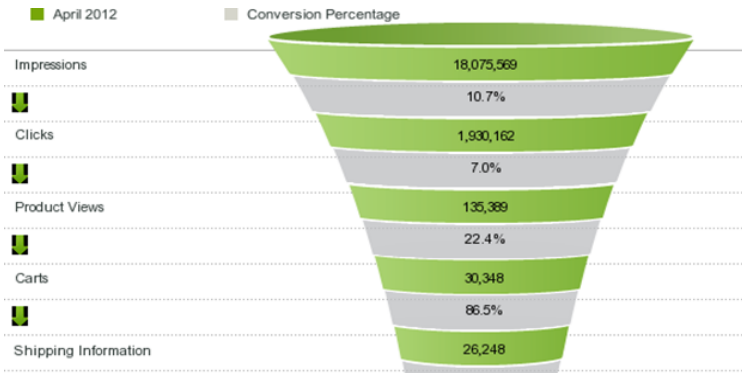
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

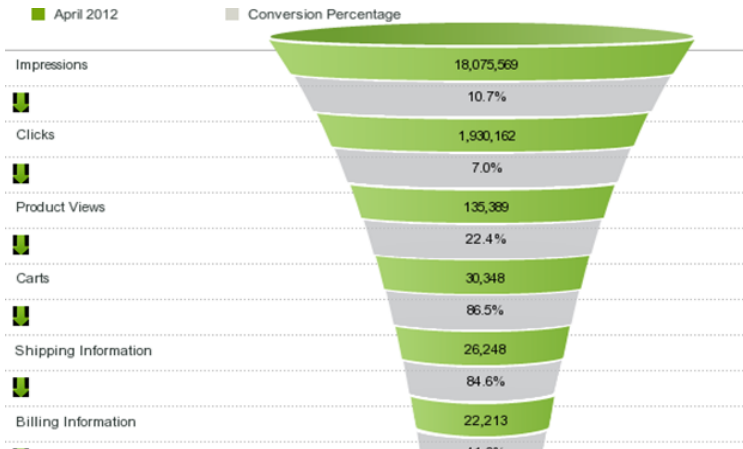
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

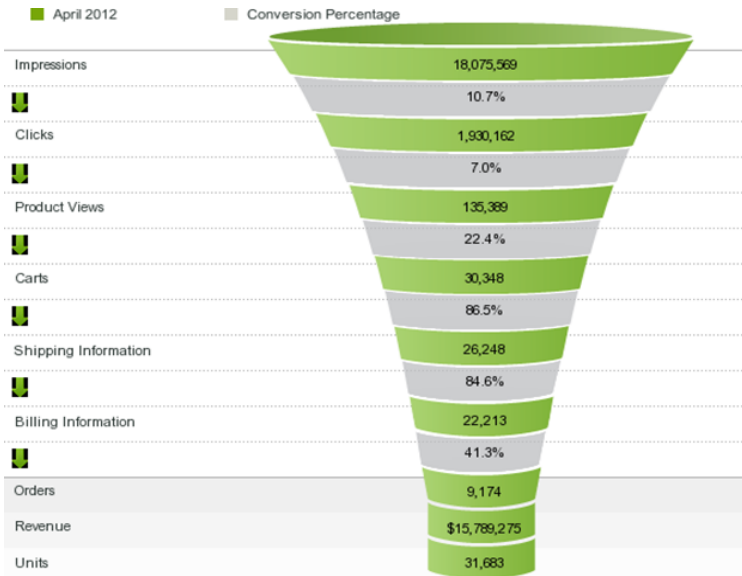
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics.



Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark
 - ▶ Queries implemented with map and reduce functions.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark
 - ▶ Queries implemented with map and reduce functions.
 - ▶ In-memory caching.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark
 - ▶ Queries implemented with map and reduce functions.
 - ▶ In-memory caching.
 - ▶ Cloudera Impala
 - ▶ Analytic Database for Apache Hadoop.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark
 - ▶ Queries implemented with map and reduce functions.
 - ▶ In-memory caching.
 - ▶ Cloudera Impala
 - ▶ Analytic Database for Apache Hadoop.
 - ▶ Google Dremel
 - ▶ Analytics of web-scale datasets.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark
 - ▶ Queries implemented with map and reduce functions.
 - ▶ In-memory caching.
 - ▶ Cloudera Impala
 - ▶ Analytic Database for Apache Hadoop.
 - ▶ Google Dremel
 - ▶ Analytics of web-scale datasets.
- ▶ We present **Spindle**, which is an early investigation of the feasibility of Apache Spark for web analytics

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

- ▶ Adobe Marketing Cloud offers web analytics for interactive data exploration.
- ▶ Terabytes of data, thousands of servers.
- ▶ Trending general-purpose distributed data processing engines.
 - ▶ Apache Spark
 - ▶ Queries implemented with map and reduce functions.
 - ▶ In-memory caching.
 - ▶ Cloudera Impala
 - ▶ Analytic Database for Apache Hadoop.
 - ▶ Google Dremel
 - ▶ Analytics of web-scale datasets.
- ▶ We present **Spindle**, which is an early investigation of the feasibility of Apache Spark for web analytics
- ▶ Goal: Low-latency query execution time.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking concurrent queries.

Scaling Spark and HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Overview.

What is Spindle?

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Overview.

What is Spindle?

```
http://server/query
```

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

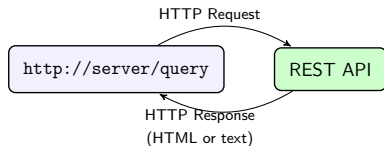
Spindle Architecture

Overview.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

What is Spindle?



Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

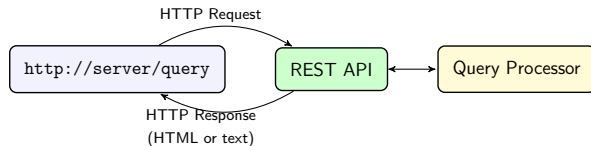
Spindle Architecture

Overview.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

What is Spindle?



Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

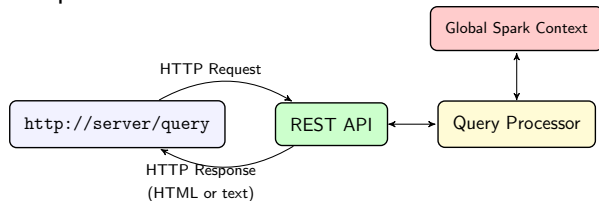
Spindle Architecture

Overview.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

What is Spindle?



Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

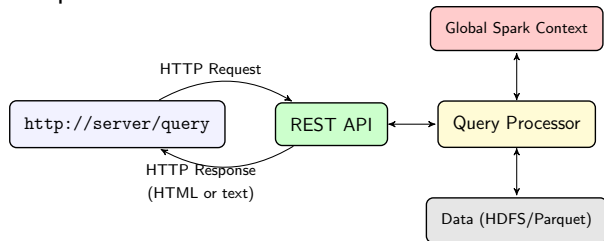
Spindle Architecture

Overview.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

What is Spindle?



Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:
 - ▶ Operates on archival data with 250 columns.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:
 - ▶ Operates on archival data with 250 columns.
 - ▶ Data is sparse and queries use <10 columns at a time.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:
 - ▶ Operates on archival data with 250 columns.
 - ▶ Data is sparse and queries use <10 columns at a time.
- ▶ Use columnar data format on distributed filesystem.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:
 - ▶ Operates on archival data with 250 columns.
 - ▶ Data is sparse and queries use <10 columns at a time.
- ▶ Use columnar data format on distributed filesystem.
- ▶ Spindle makes tuning parameters easy.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:
 - ▶ Operates on archival data with 250 columns.
 - ▶ Data is sparse and queries use <10 columns at a time.
- ▶ Use columnar data format on distributed filesystem.
- ▶ Spindle makes tuning parameters easy.
 - ▶ Intermediate data partitioning

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Features.

- ▶ Data format challenges:
 - ▶ Operates on archival data with 250 columns.
 - ▶ Data is sparse and queries use <10 columns at a time.
- ▶ Use columnar data format on distributed filesystem.
- ▶ Spindle makes tuning parameters easy.
 - ▶ Intermediate data partitioning
 - ▶ Caching

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Queries.

- ▶ Experimental setup: Representative set of analytics queries.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Queries.

- ▶ Experimental setup: Representative set of analytics queries.

Shorthand	Name
Q0	Pageviews
Q1	Revenue
Q2	RevenueFromTopReferringDomains
Q3	RevenueFromTopReferringDomainsFirstVisitGoogle
Q4	TopPages
Q5	TopPagesByBrowser
Q6	TopPagesByPreviousTopPages
Q7	TopReferringDomains

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Queries.

- ▶ Queries use a small columnar subset.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Spindle Architecture

Queries.

- ▶ Queries use a small columnar subset.

	Q0	Q1	Q2	Q3	Q4	Q5	Q6	Q7
post_pagename	x				x	x	x	
user_agent						x		
visit_referrer			x	x				
post_visit_high			x	x			x	x
post_visit_low			x	x			x	x
visit_num			x	x			x	x
visit_referrer								x
hit_time_gmt							x	
post_purchaseid		x	x	x				
post_product_list		x	x	x				
first_hit_referrer				x				

Columns

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking concurrent queries.

Scaling Spark and HDFS workers.

Future Work

Conclusions

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Caching.

- ▶ Six cluster nodes (32 GB memory each), Spark and HDFS on each.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Caching.

- ▶ Six cluster nodes (32 GB memory each), Spark and HDFS on each.
- ▶ 13.1GB of data, 1 week, 1 customer.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Caching.

- ▶ Six cluster nodes (32 GB memory each), Spark and HDFS on each.
- ▶ 13.1GB of data, 1 week, 1 customer.
- ▶ **Question:** How does caching in-memory improve performance?

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

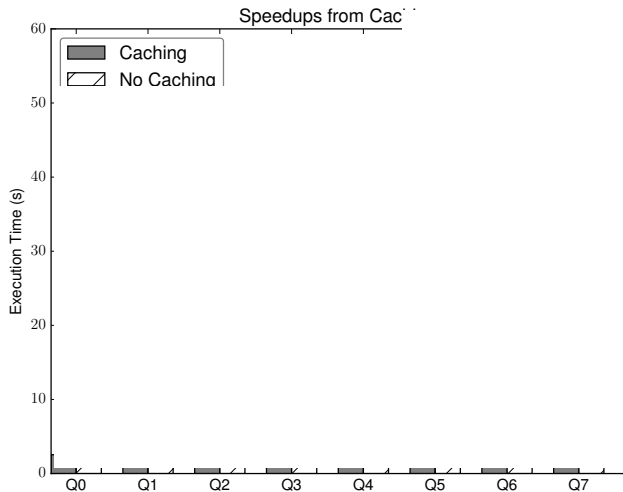
Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions



Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

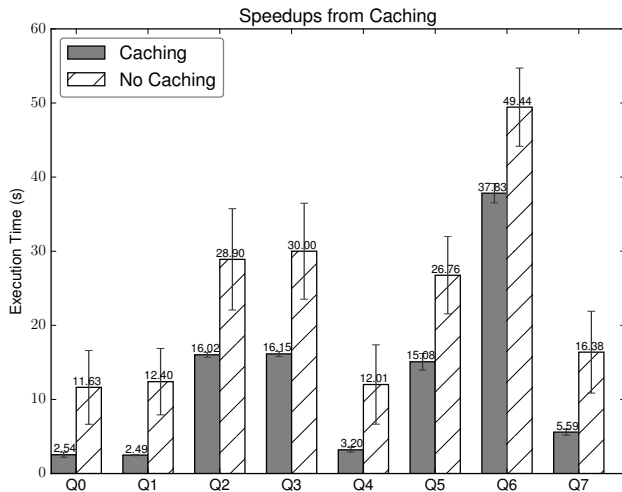
Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions



Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking

concurrent

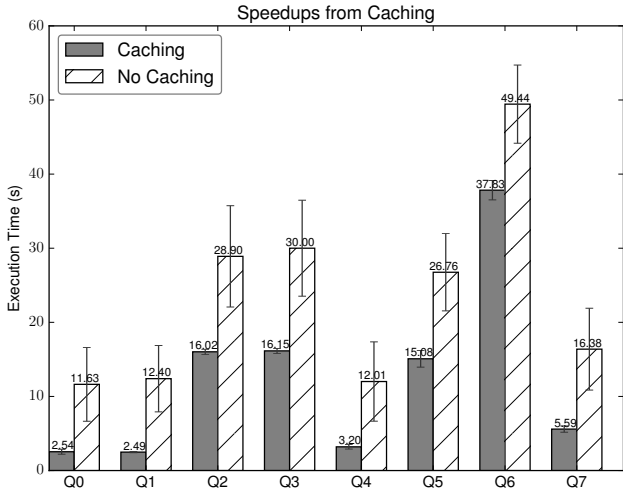
queries.

Scaling Spark and

HDFS workers.

Future Work

Conclusions



Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.

Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

- Caching helps, but what else can be done to lower query execution times?

Empirical Results

Data partitioning.

- ▶ Partitions are groups of data executed in a batch.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Data partitioning.

- ▶ Partitions are groups of data executed in a batch.
- ▶ Partitions can be executed concurrently.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Data partitioning.

- ▶ Partitions are groups of data executed in a batch.
- ▶ Partitions can be executed concurrently.
- ▶ Not clear how to partition the intermediate data.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Data partitioning.

- ▶ Partitions are groups of data executed in a batch.
- ▶ Partitions can be executed concurrently.
- ▶ Not clear how to partition the intermediate data.
 - ▶ Too small: Partition management overhead.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Data partitioning.

- ▶ Partitions are groups of data executed in a batch.
- ▶ Partitions can be executed concurrently.
- ▶ Not clear how to partition the intermediate data.
 - ▶ Too small: Partition management overhead.
 - ▶ Too large: Data is processed in serial.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

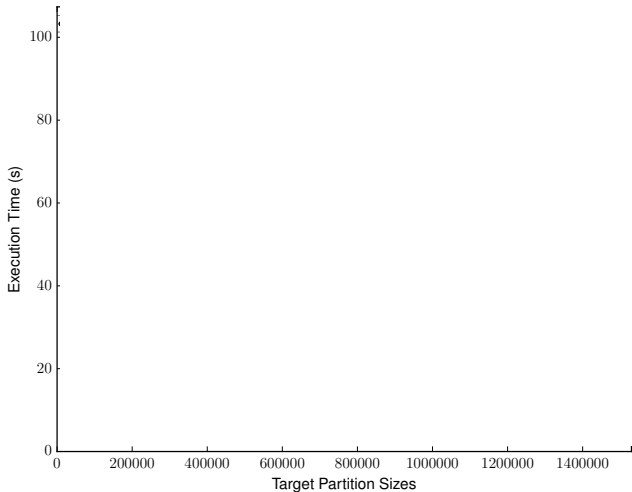
Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions



Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.

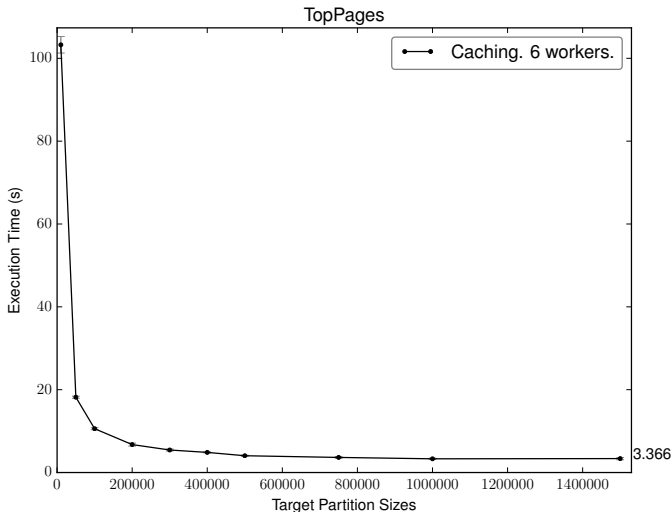
Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions



► Targeting 1.5M items in each partition is reasonable.

Empirical Results

Benchmarking concurrent queries.

- ▶ How much will Spindle's performance degrade if multiple users are utilizing it at the same time?

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
**Benchmarking
concurrent
queries.**
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Benchmarking concurrent queries.

- ▶ How much will Spindle's performance degrade if multiple users are utilizing it at the same time?
- ▶ Concurrently call the same query on the same data.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
**Benchmarking
concurrent
queries.**
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Empirical Results

Benchmarking concurrent queries.

- ▶ How much will Spindle's performance degrade if multiple users are utilizing it at the same time?
- ▶ Concurrently call the same query on the same data.
- ▶ Average execution times.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
**Benchmarking
concurrent
queries.**
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

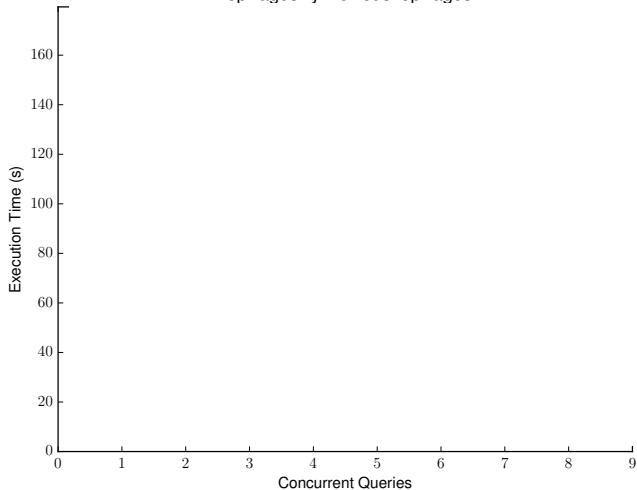
**Benchmarking
concurrent
queries.**

Scaling Spark and
HDFS workers.

Future Work

Conclusions

TopPagesByPreviousTopPages



Motivation

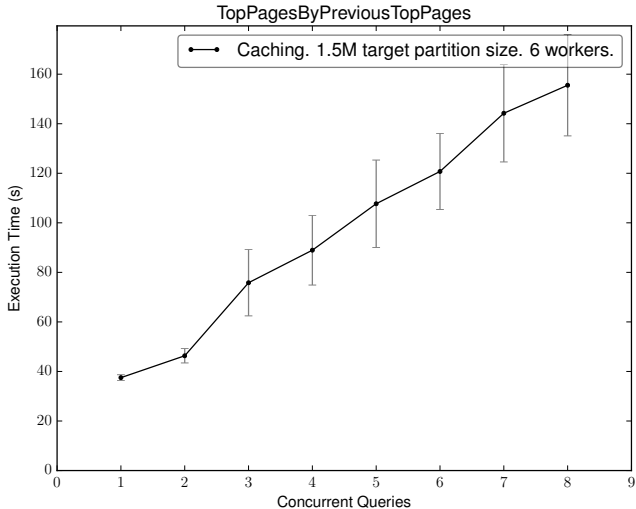
Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
**Benchmarking
concurrent
queries.**
Scaling Spark and
HDFS workers.

Future Work

Conclusions



- Performance better than serializing concurrent requests, but can be improved.

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking

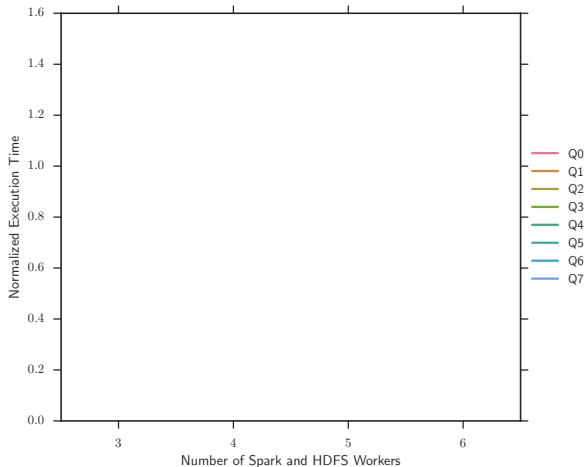
concurrent

queries.

**Scaling Spark and
HDFS workers.**

Future Work

Conclusions



Motivation

Spindle
Architecture
Overview.
Features.
Queries.

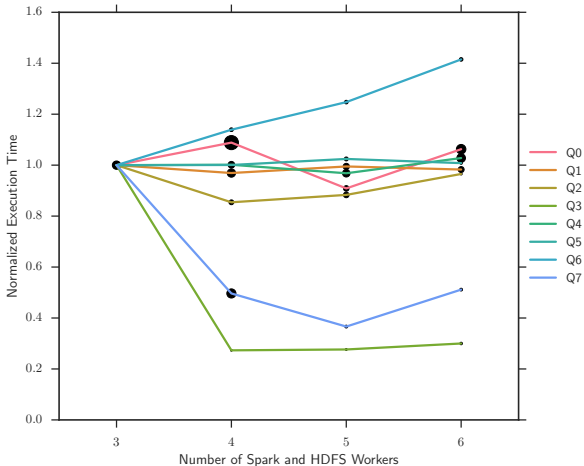
Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.

**Scaling Spark and
HDFS workers.**

Future Work

Conclusions



- Further profiling is needed to improve performance as increasing the number of workers.

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking concurrent queries.

Scaling Spark and HDFS workers.

Future Work

Conclusions

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.
 - ▶ Online/Dynamically

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.
 - ▶ Online/Dynamically
 - ▶ Offline

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.
 - ▶ Online/Dynamically
 - ▶ Offline
- ▶ Results caching for identical queries.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.
 - ▶ Online/Dynamically
 - ▶ Offline
- ▶ Results caching for identical queries.
- ▶ Data preprocessing to remove redundant computations.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.
 - ▶ Online/Dynamically
 - ▶ Offline
- ▶ Results caching for identical queries.
- ▶ Data preprocessing to remove redundant computations.
- ▶ Distributed filesystem caching with Tachyon.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Future Work

- ▶ Lowering query execution time.
 - ▶ Goal: Sub-second.
- ▶ Automatically tuning parameter exploration space for a given workload.
 - ▶ Online/Dynamically
 - ▶ Offline
- ▶ Results caching for identical queries.
- ▶ Data preprocessing to remove redundant computations.
- ▶ Distributed filesystem caching with Tachyon.
- ▶ Optimized query generation with SparkSQL.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Motivation

Spindle Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking concurrent queries.

Scaling Spark and HDFS workers.

Future Work

Conclusions

Motivation

Spindle
Architecture

Overview.

Features.

Queries.

Empirical Results

Caching.

Data partitioning.

Benchmarking
concurrent
queries.

Scaling Spark and
HDFS workers.

Future Work

Conclusions

Conclusions

- ▶ We present **Spindle**.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture

Overview.
Features.
Queries.

Empirical Results

Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Conclusions

- ▶ We present **Spindle**.
 - ▶ **Open-source** prototype analytics processing engine.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Conclusions

- ▶ We present **Spindle**.
 - ▶ **Open-source** prototype analytics processing engine.
 - ▶ Sample set of web analytics queries.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Conclusions

- ▶ We present **Spindle**.
 - ▶ **Open-source** prototype analytics processing engine.
 - ▶ Sample set of web analytics queries.
 - ▶ Interface for parameter tuning.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Conclusions

- ▶ We present **Spindle**.
 - ▶ **Open-source** prototype analytics processing engine.
 - ▶ Sample set of web analytics queries.
 - ▶ Interface for parameter tuning.

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions

Conclusions

- ▶ We present **Spindle**.
 - ▶ **Open-source** prototype analytics processing engine.
 - ▶ Sample set of web analytics queries.
 - ▶ Interface for parameter tuning.

Spindle Project		http://github.com/adobe-research/spindle
Demo		http://adobe-research.github.io/spindle/
Brandon Amos		http://github.com/bamos
David Tompkins		http://github.com/DavidTompkins

Spindle,
CloudCom 2014

Amos and
Tompkins,
Adobe Research

Motivation

Spindle
Architecture
Overview.
Features.
Queries.

Empirical Results
Caching.
Data partitioning.
Benchmarking
concurrent
queries.
Scaling Spark and
HDFS workers.

Future Work

Conclusions