DRIVING INNOVATION THROUGH DATA

ACCELERATING BIG DATA APPLICATION DEVELOPMENT WITH CASCADING

Supreet Oberoi
VP Field Engineering, Concurrent Inc

CONCURRENT

GET TO KNOW CONCURRENT

CONCURRENT

Founded: *2008*
HQ: *San Francisco, CA*

CEO: *Gary Nakamura*
CTO, Founder: *Chris Wensel*

www.concurrentinc.com

Leader in Application Infrastructure for Big Data

- Building enterprise software to simplify Big Data application development and management

Products and Technology

- ***CASCADING***

  Open Source - The most widely used application infrastructure for building Big Data apps with over 175,000 downloads each month

- ***DRIVEN***

  Enterprise data application management for Big Data apps

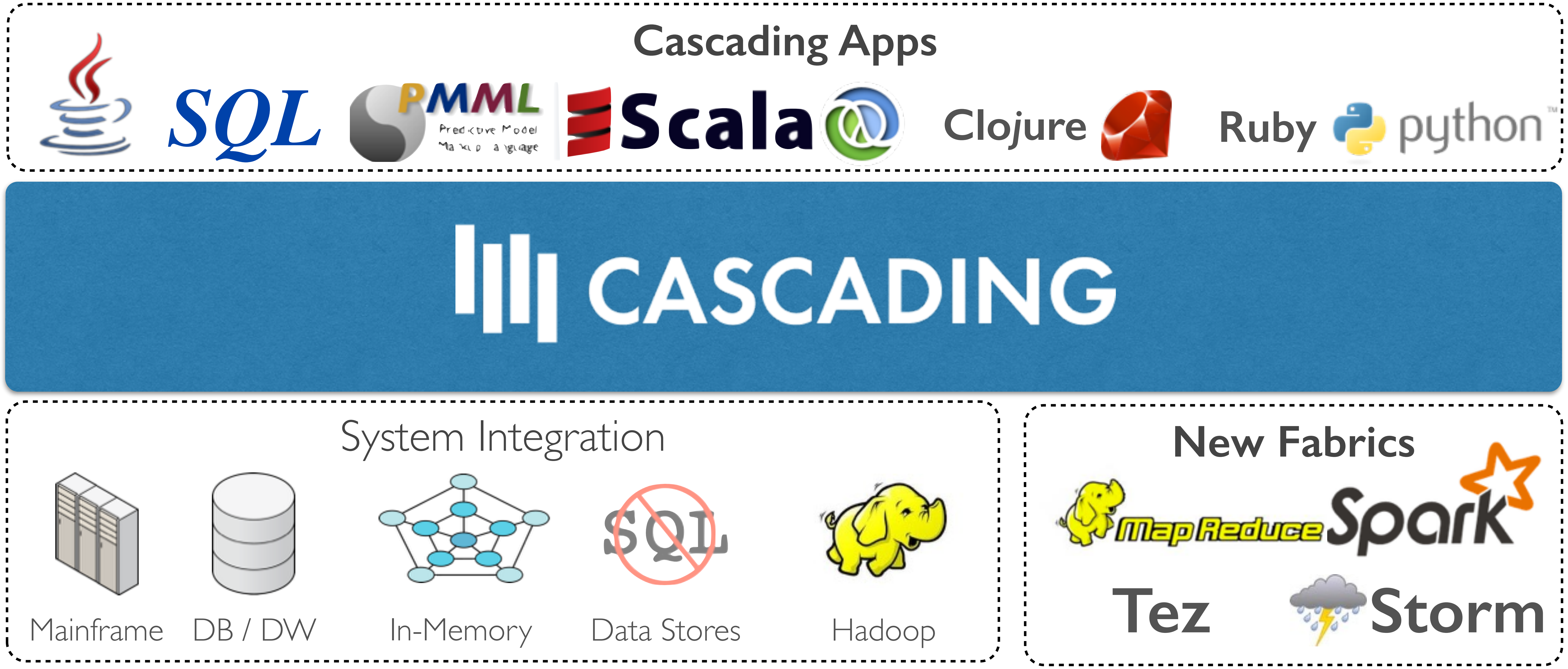
Proven — Simple, Reliable, Robust

- Thousands of enterprises rely on Concurrent to provide their data application infrastructure.

ENTERPRISE NEEDS FOR DATA APP INFRASTRUCTURE

- Need reliable, reusable tooling to quickly build and consistently deliver data products
- Need the degrees of freedom to solve problems ranging from simple to complex with existing skill sets
- Need the flexibility to easily adapt an application to meet business needs (latency, scale, SLA), without having to rewrite the application
- Need operational visibility for entire data application lifecycle

CASCADING - DE-FACTO FRAMEWORK FOR DATA APPS



Cascading Apps

SQL PMML Scala Clojure Ruby python

CASCADING

System Integration

Mainframe DB / DW In-Memory Data Stores Hadoop

New Fabrics

MapReduce Spark Tez Storm

- Standard for enterprise data app development
- Your programming language of choice
- Cascading applications that run on MapReduce will also run on Apache Tez, Spark, Storm, and …

WORD COUNT EXAMPLE WITH CASCADING

```
  String docPath = args[ 0 ];
  String wcPath = args[ 1 ];
  Properties properties = new Properties();
  AppProps.setApplicationJarClass( properties, Main.class );
  HadoopFlowConnector flowConnector = new HadoopFlowConnector( properties );
```

configuration

```
 // create source and sink taps
 Tap docTap = new Hfs( new TextDelimited( true, "\t" ), docPath );
 Tap wcTap = new Hfs( new TextDelimited( true, "\t" ), wcPath );
```

integration

```
// specify a regex to split "document" text lines into token stream
Fields token = new Fields( "token" );
Fields text = new Fields( "text" );
RegexSplitGenerator splitter = new RegexSplitGenerator( token, "[ \\[\\]\\(\\),.]" );
// only returns "token"
Pipe docPipe = new Each( "token", text, splitter, Fields.RESULTS );
// determine the word counts
Pipe wcPipe = new Pipe( "wc", docPipe );
wcPipe = new GroupBy( wcPipe, token );
wcPipe = new Every( wcPipe, Fields.ALL, new Count(), Fields.ALL );
```

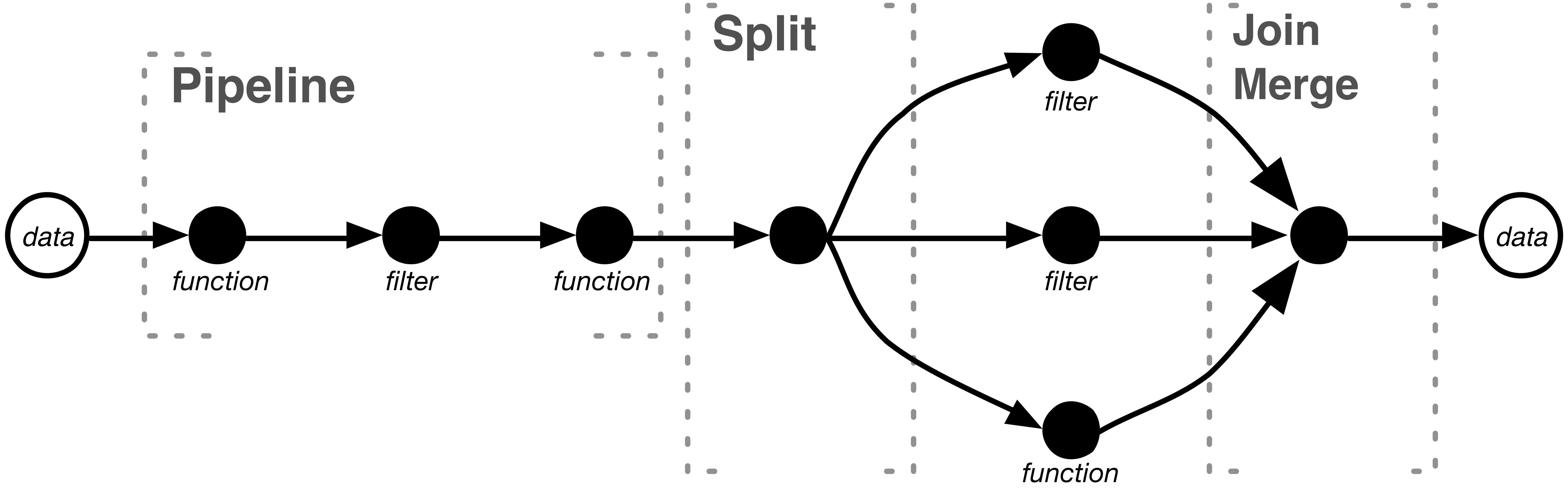
processing

```
// connect the taps, pipes, etc., into a flow definition
FlowDef flowDef = FlowDef.flowDef().setName( "wc" )
 .addSource( docPipe, docTap )
 .addTailSink( wcPipe, wcTap );
// create the Flow
Flow wcFlow = flowConnector.connect( flowDef ); // <<-- Unit of Work
wcFlow.complete();                              // <<-- Runs jobs on Cluster
```

scheduling

SOME COMMON PATTERNS

- Functions
- Filters
- Joins
  - Inner / Outer / Mixed
  - Asymmetrical / Symmetrical
- Merge (Union)
- Grouping
  - Secondary Sorting
  - Unique (Distinct)
- Aggregations
  - Count, Average, etc

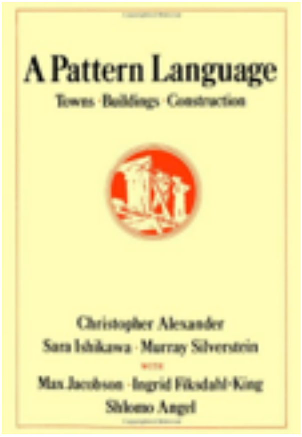


Topology

CONCURRENT

PLUMBING METAPHOR FOR BUILDING DATA FLOWS

- The Cascading processing model is based on a metaphor of flows based on patterns



A Pattern Language

Towns · Buildings · Construction

Christopher Alexander
Sara Ishikawa · Murray Silverstein
with
Max Jacobson · Ingrid Fiksdahl-King
Shlomo Angel

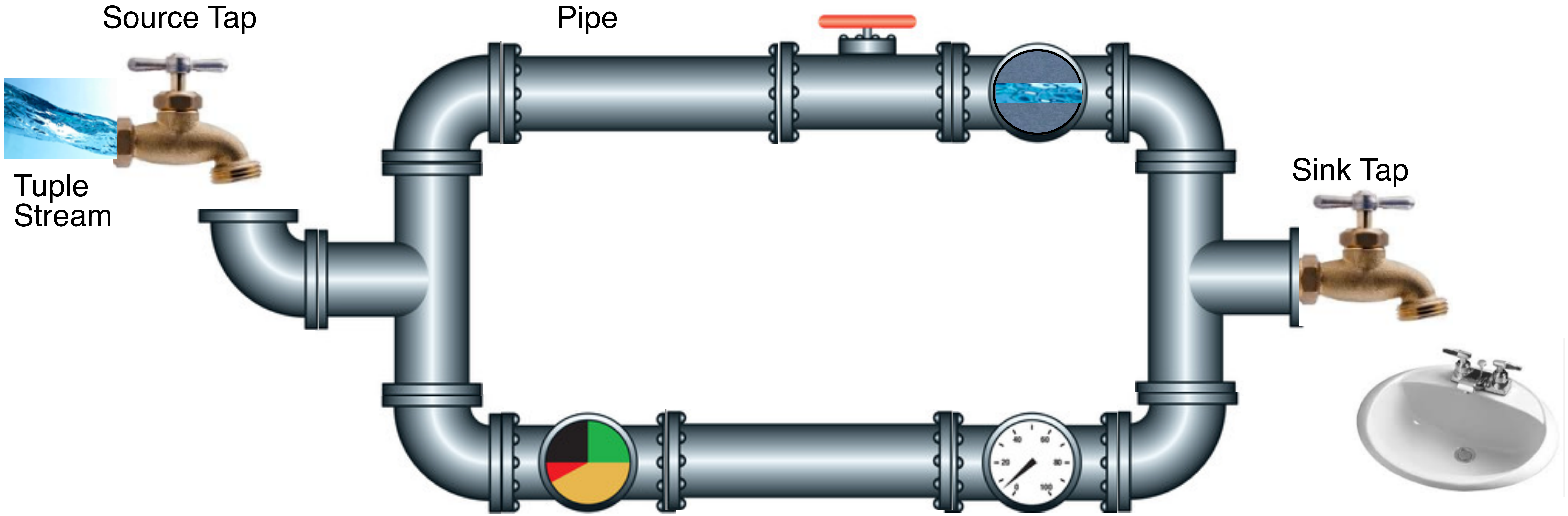
Design Patterns

Elements of Reusable Object-Oriented Software

Erich Gamma
Richard Helm
Ralph Johnson
John Vlissides

Foreword by Grady Booch

ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES



CASCADING PROCESSING MODEL TERMINOLOGY

Tuple Stream	Series of tuples (data record)
Fields	Representation of the Tuple Stream, used in operations
Pipe	Applies operations to tuples or groups of tuples
Branch	Pipes linked together under a common Pipe name
Pipe Assembly	An interconnected set of pipe branches
Tap	Source or sink for data
Flow	Pipe assembly with taps
Cascade	Multiple flows grouped together & executed as a single process

TUPLE STREAM

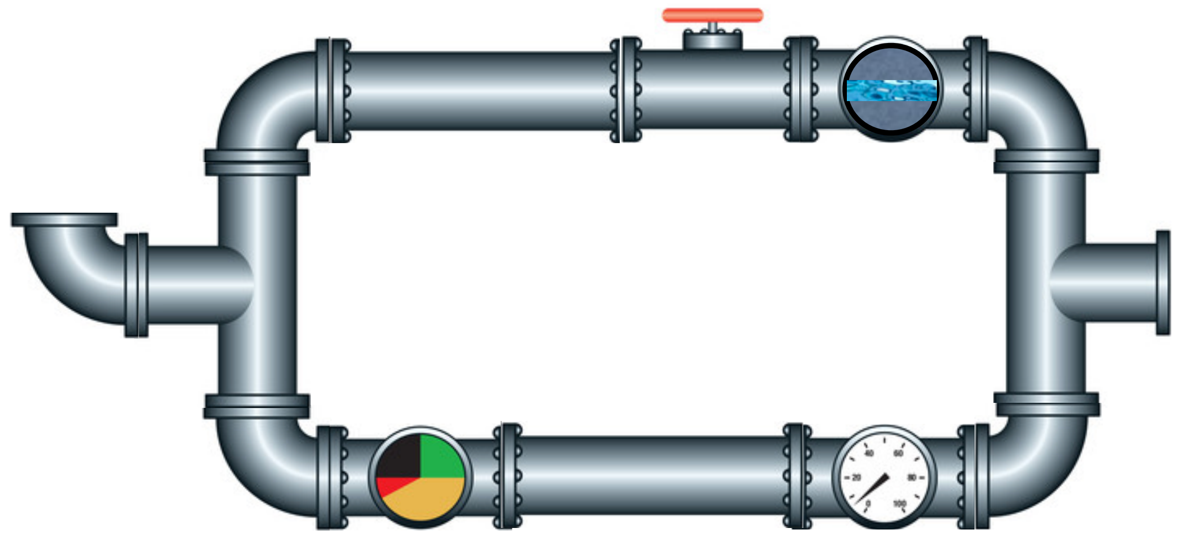
- A Tuple represents a set of values.
- Consider a Tuple the same as a database record where every value is a column in that table.
- A "tuple stream" is a set of Tuple instances passed consecutively through a Pipe assembly.

PIPES CAN BE CHAINED TO PERFORM COMPLEX OPERATIONS

- Pipes control the flow of data applying operations to each Tuple or groups of Tuples.
- Pipes work on fields of one or more tuples.
- Pipes allow you to manage a data flow such as doing:
  - Grouping
  - Joining
  - Filtering
  - Buffering
  - Aggregating

PIPES CAN BE BRANCHED AND MERGED

- Pipe Assemblies are an interconnected set of pipe branches modeled as a DAG (Directed Acyclic Graph)
- Pipe Assemblies can consist of splits and/or merges.
- Pipe assemblies are specified independently of the data source they are to process.
- For a pipe assembly to be executed, it must be bound to data sources and sinks (which becomes a flow)



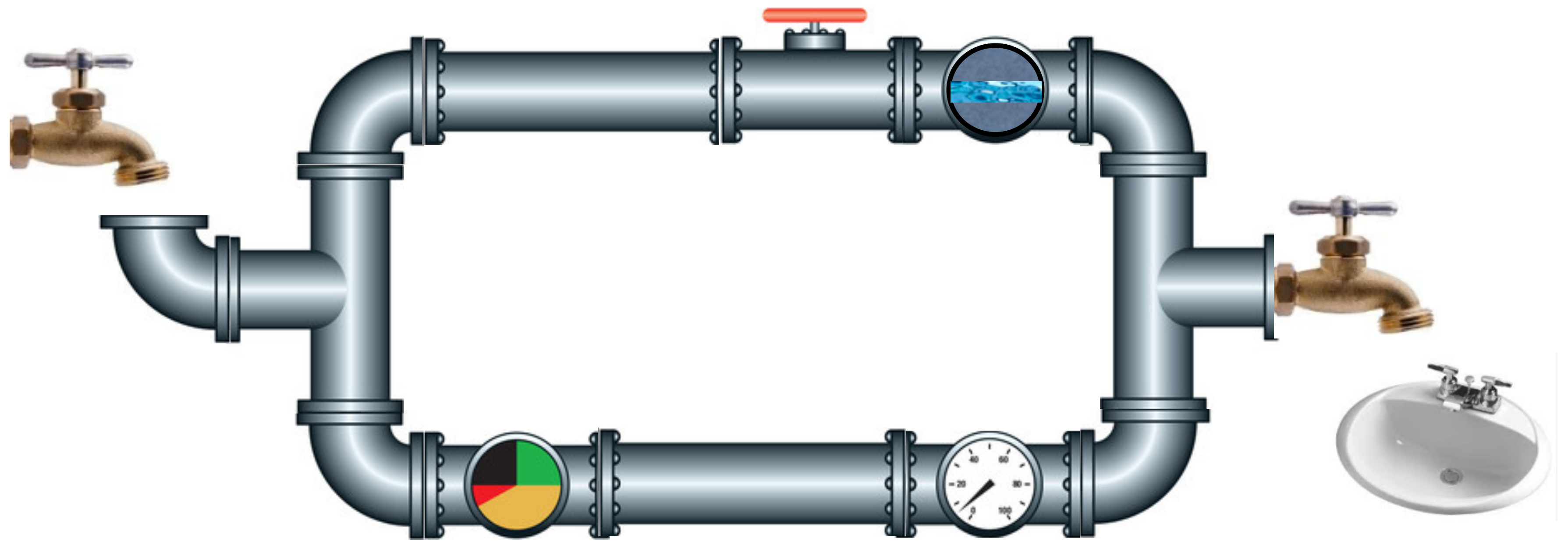
DAG: collection of vertices and directed edges, each edge connecting one vertex to another, such that there is no way to start at some vertex v and follow a sequence of edges that eventually loops back to v again.

TAPS ABSTRACT INTEGRATION TO THIRD-PARTY SYSTEMS

- Taps provide the ability to read and write data.
- Taps can be shared between flows and can be restricted to being either sources or sinks.
- Taps can be set up to have the actual file identifiers determined when they run.
- Examples of Taps are:
  - File on the local file system
  - File on a Hadoop distributed file system
  - File on Amazon S3

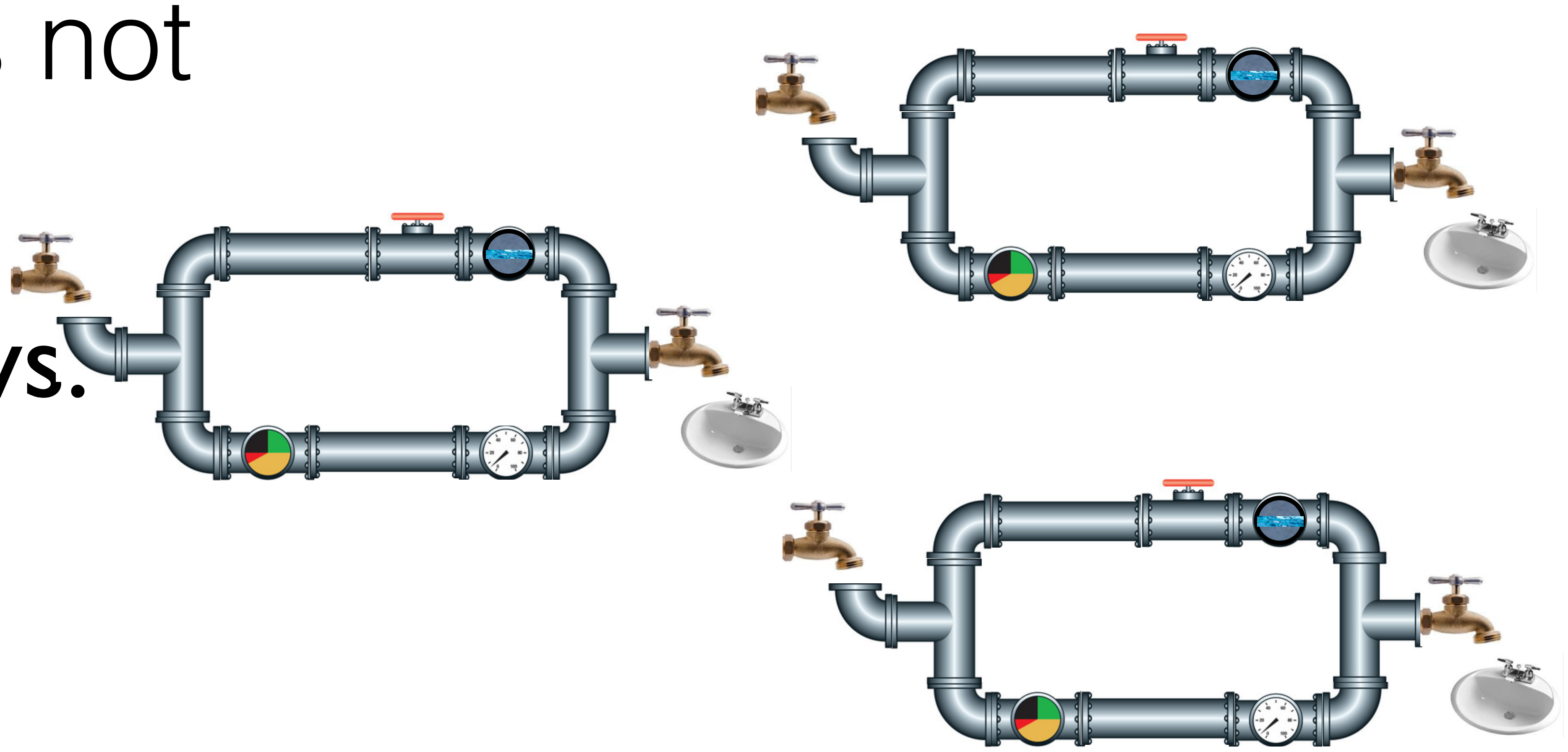
FLOWS CONNECT IT ALL TOGETHER FOR EXECUTION

- Flows consist of pipe assemblies with data sources and sinks
- Flows contain one or more data sources, a DAG (Directed Acyclic Graph) of pipes, and one or more data sinks.
- Flows are designed to be re-useable units of work.
- Flows show the business and programming process.
- A flow is a basic unit of work of arbitrary size.



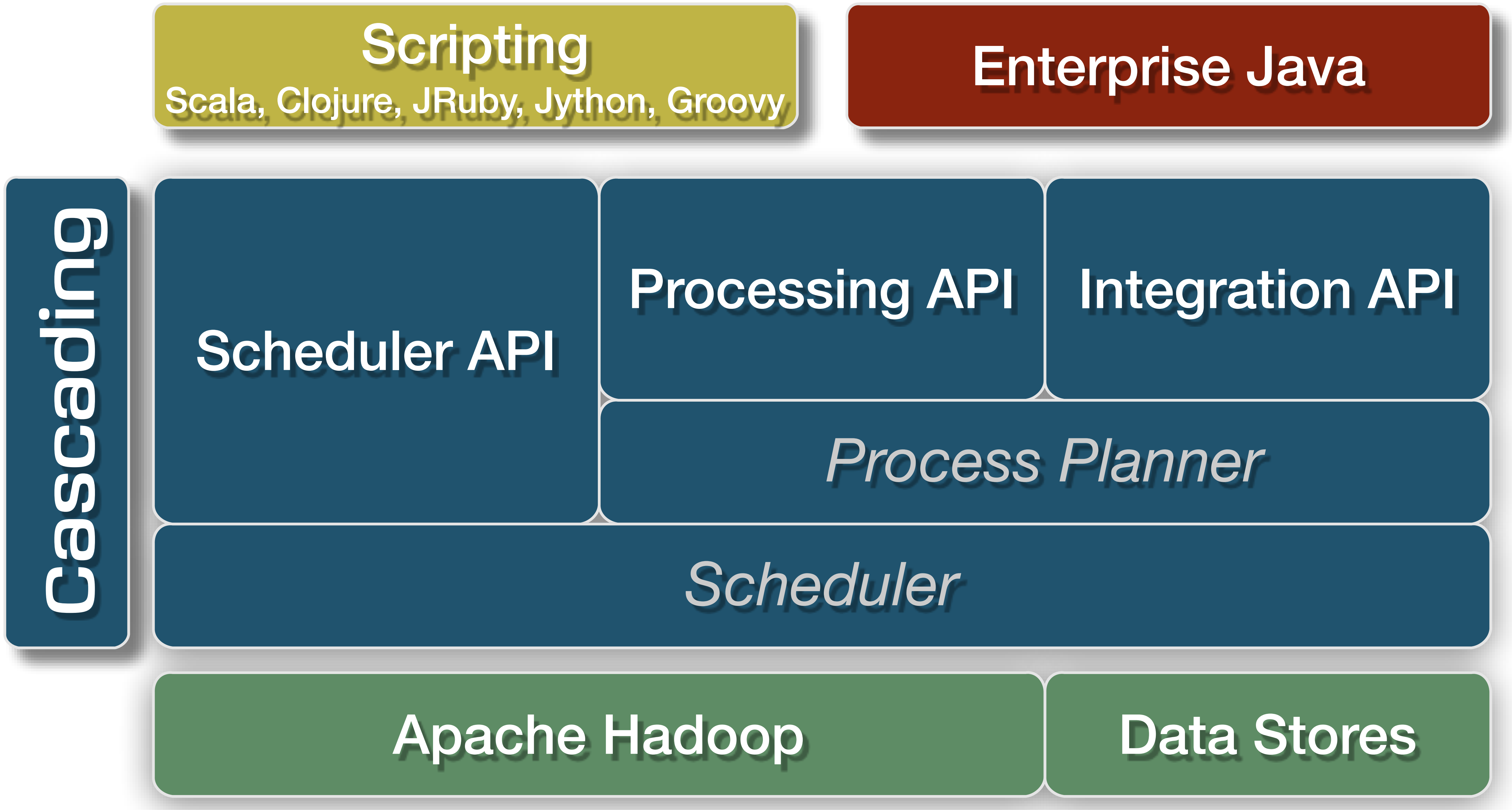
FLOWS CAN BE CONNECTED INTO A CASCADE

- Cascade joins together multiple flows.
- Use Cascade if there are dependencies among the Flows:
  - Cascade will cause a flow to not be executed until all of its data dependencies are satisfied.
  - A cascade can determine that a Flow does not need to run.
- A CascadeConnector makes a Cascade from Flows.

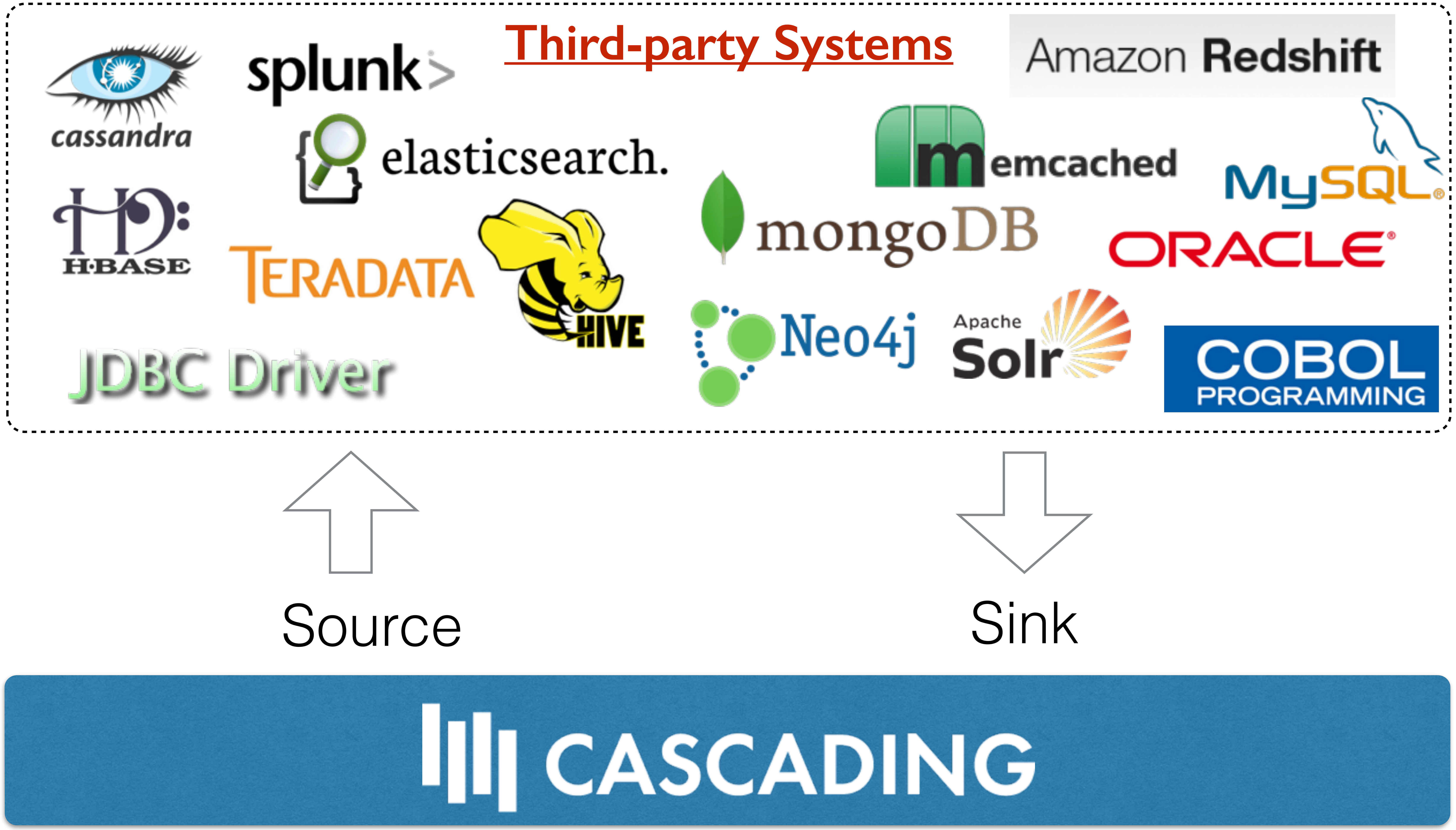


CASCADING RUNTIME FRAMEWORK ABSTRACTS INTEGRATION & COMPUTE FABRIC

- Java API
- Separates business logic from integration
- Testable at every lifecycle stage
- Works with any JVM language
- Many integration adapters



CASCADING - INTEGRATION WITH EXTERNAL SYSTEMS



Third-party Systems

Source

Sink

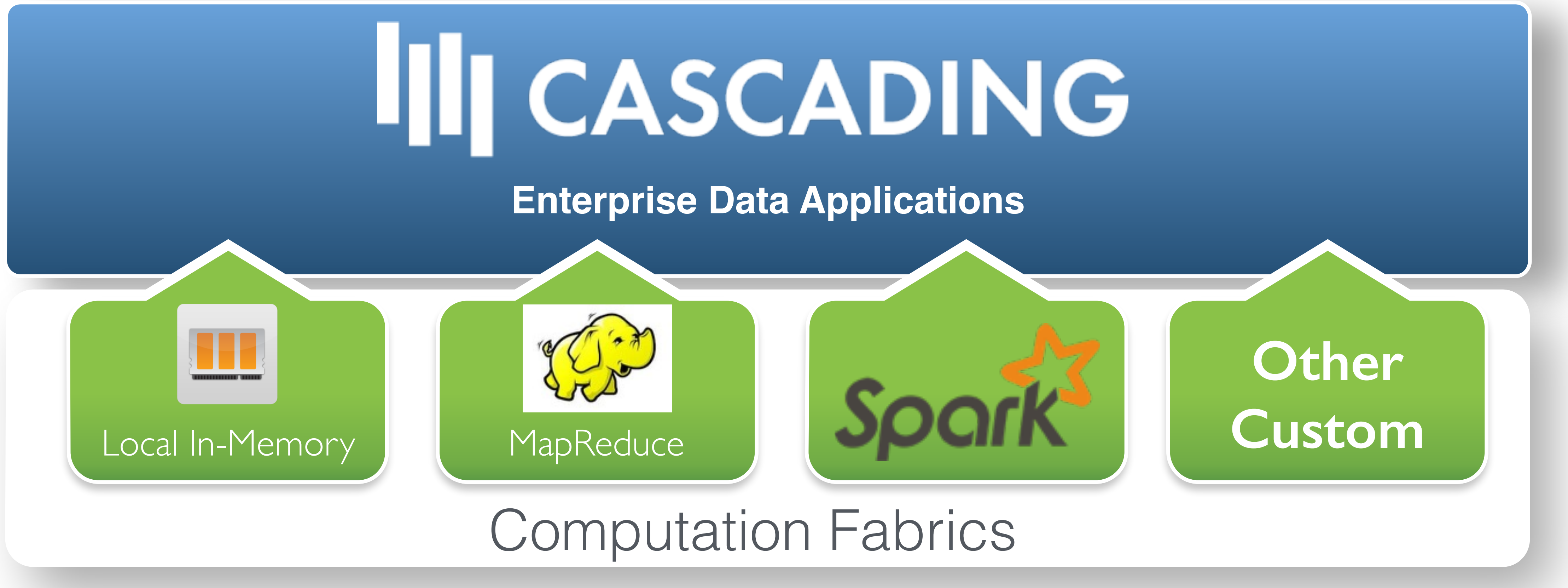
CASCADING

http://www.cascading.org/extensions/

CASCADING - APP PORTABILITY

“Write once and deploy on your fabric of choice.”

- The Innovation — Cascading allows for data apps to execute on existing and emerging fabrics through its new customizable query planner.
- Cascading 3.0 supports — Local In-Memory, Apache MapReduce and Apache Tez. 1H 2015 - Apache Spark and Apache Storm
- Flexibility to meet changing business needs



THE STANDARD FOR DATA APPLICATION DEVELOPMENT

CASCADING

Proven application development framework for building data apps

www.cascading.org

Application platform that addresses:

Build data apps that are scale-free

Design principals ensure best practices at any scale

Systems Integration

Hadoop never lives alone. Easily integrate to existing systems

Application Portability

Write once, then run on different computation fabrics

Staffing Bottleneck

Use existing Java,Scala, SQL, modeling skill sets

Test-Driven Development

Efficiently test code and process local files before deploying on a cluster

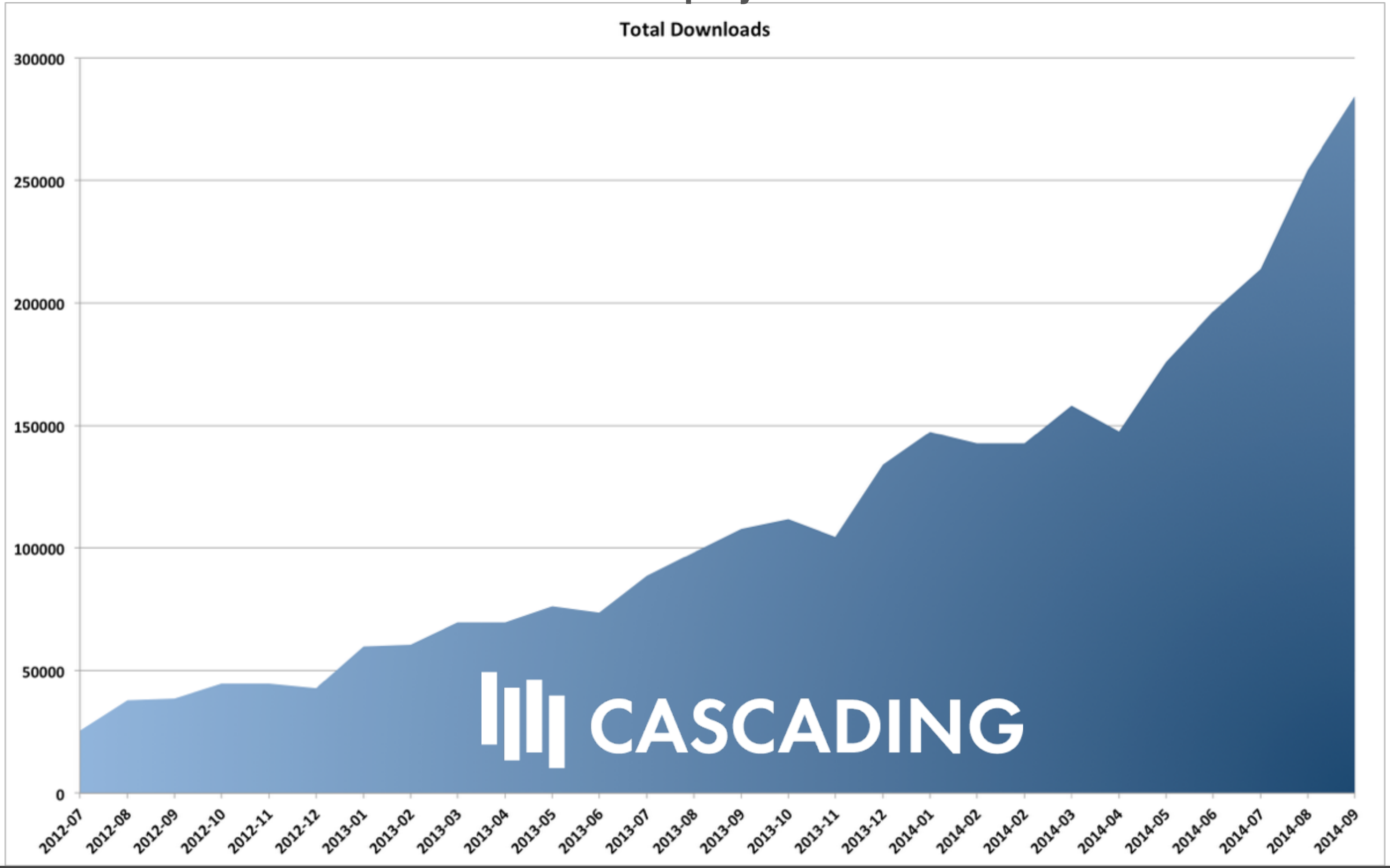
Operational Complexity

Simple - Package up into one jar and hand to operations

STRONG ORGANIC GROWTH

280K+ downloads / month

7000+ Deployments



CASCADING DATA APPLICATIONS

CASCADING

Enterprise IT

Extract Transform Load
Log File Analysis
Systems Integration
Operations Analysis

Corporate Apps

HR Analytics
Employee Behavioral Analysis
Customer Support | eCRM
Business Reporting

Telecom

Data processing of Open Data
Geospatial Indexing
Consumer Mobile Apps
Location based services

Marketing / Retail

Mobile, Social, Search Analytics
Funnel Analysis
Revenue Attribution
Customer Experiments
Ad Optimization
Retail Recommenders

Consumer / Entertainment

Music Recommendation
Comparison Shopping
Restaurant Rankings
Real Estate
Rental Listings
Travel Search & Forecast

Finance

Fraud and Anomaly Detection
Fraud Experiments
Customer Analytics
Insurance Risk Metric

Health / Biotech

Aggregate Metrics For Govt
Person Biometrics
Veterinary Diagnostics
Next-Gen Genomics
Argonomics
Environmental Maps

BUSINESSES DEPEND ON US

- Cascading Java API
- Data normalization and cleansing of search and click-through logs for use by analytics tools, Hive analysts
- Easy to operationalize heavy lifting of data in one framework

BUSINESSES DEPEND ON US

CASCADING

THE CLIMATE CORPORATION

- Cascalog (Clojure)
- Weather pattern modeling to protect growers against loss
- ETL against 20+ datasets daily
- Machine learning to create models
- Purchased by Monsanto for $930M US

BUSINESSES DEPEND ON US

CASCADING

TWITTER

- Scalding (Scala)
- Makes complex analysis of very large data sets simple
- Machine learning, linear algebra to improve
- 30,000 jobs a day — this works @ scale
- Ad quality (matching users and ad effectiveness)

CASCADING DEPLOYMENTS

CASCADING

ebay
NOKIA
Square
yp
Etsy
THE CLIMATE CORPORATION
KOHL'S
Pitney Bowes
twitter
WILLIAMS-SONOMA
airbnb
bloomreach GET FOUND.
ENTAGEN
SOUNDCLOUD
MONSANTO
Telefonica
TELENAV
verizon
trulia
tresata
ak MAKING MEDIA ACCOUNTABLE
Nextag
sky
KLOUT
Citysearch
wibidata
eHarmony
LexisNexis
factual.
SK planet
Switch
EARLY WARNING
Neiman Marcus
CBS Interactive
hp
OpenTable
BlackBerry
SOLUSI247 24 hours 7 days integrated ICT solution
VALO RESEARCH TRADING
DELTA
BEST BUY
LinkedIn
DataSong
bluekai
yelp
Deutsche Bank
PivotLink
yieldbot
VISA
GENOME Institute of Singapore
shopify
ORBITZ
TREND MICRO
TWO SIGMA
StumbleUpon
lumosity
ORDINA
LiveRamp
SAMSUNG
stripe
appnexus
HARVARD School of Public Health
LIVEPERSON
FREEVERSE
Cinnober
CITYGRID
VIADEN
DEEP VALUE
Pinterest
IFS INTERNATIONAL FUND SERVICES A State Street Company
METAMARKETS
Healthline Connect to Better Health
VisualDNA
Spotify
RADIUS
criteo
BandPage
UnitedHealthcare

BROAD SUPPORT

Hadoop ecosystem supports Cascading

cloudera®

Hortonworks

MAPR™ TECHNOLOGIES

amazon webservices™

PIVOTAL HD

Microsoft

rackspace® the open cloud company

DATABRICKS

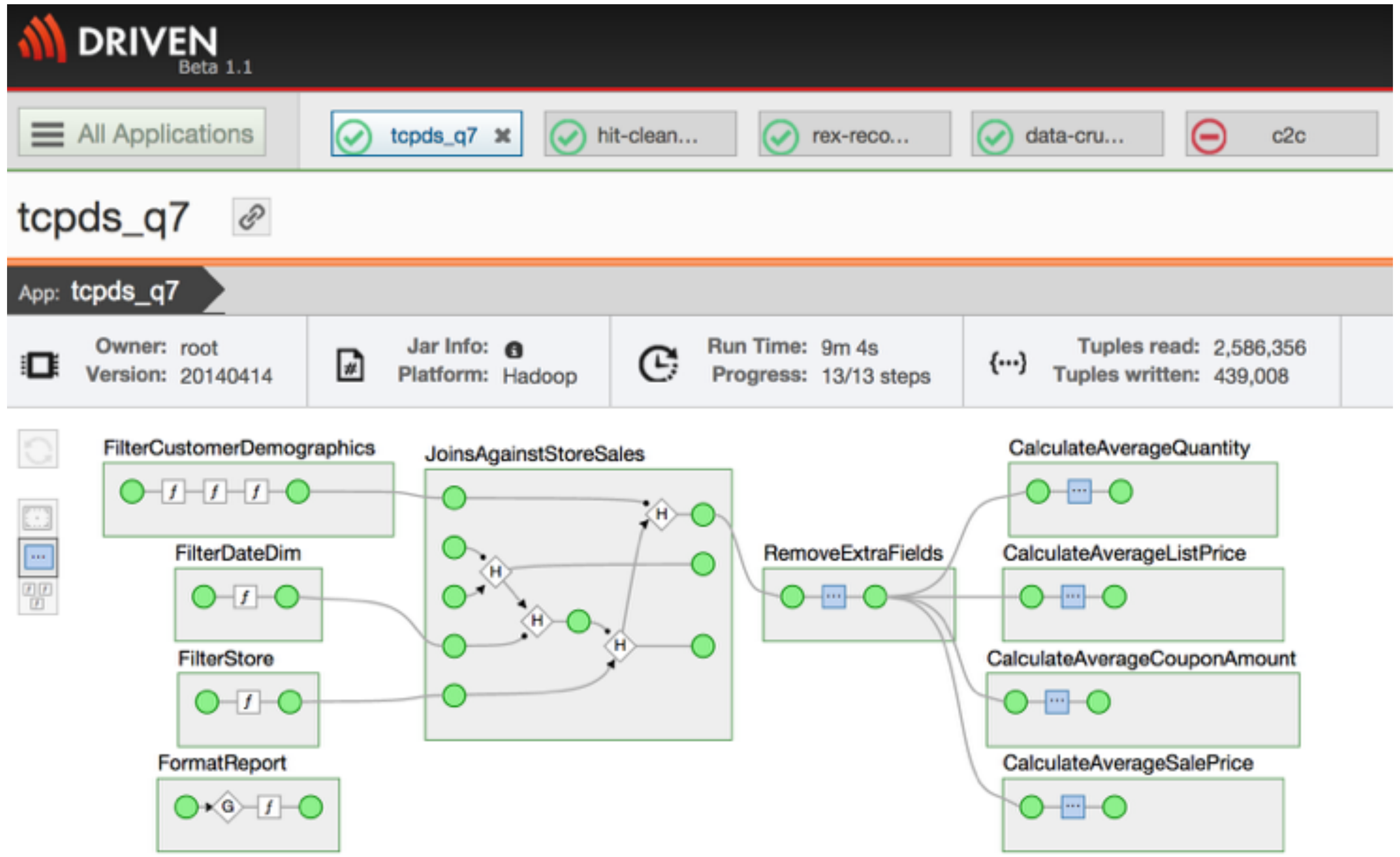
Qubole

altiscale

IBM®

OPERATIONAL EXCELLENCE WITH DRIVEN

Visibility from Development to Production



Owner: root
Version: 20140414

Jar Info:
Platform: Hadoop

Run Time: 9m 4s
Progress: 13/13 steps

Tuples read: 2,586,356
Tuples written: 439,008

Development — Building and Testing

- Design & Development
- Debugging
- Tuning

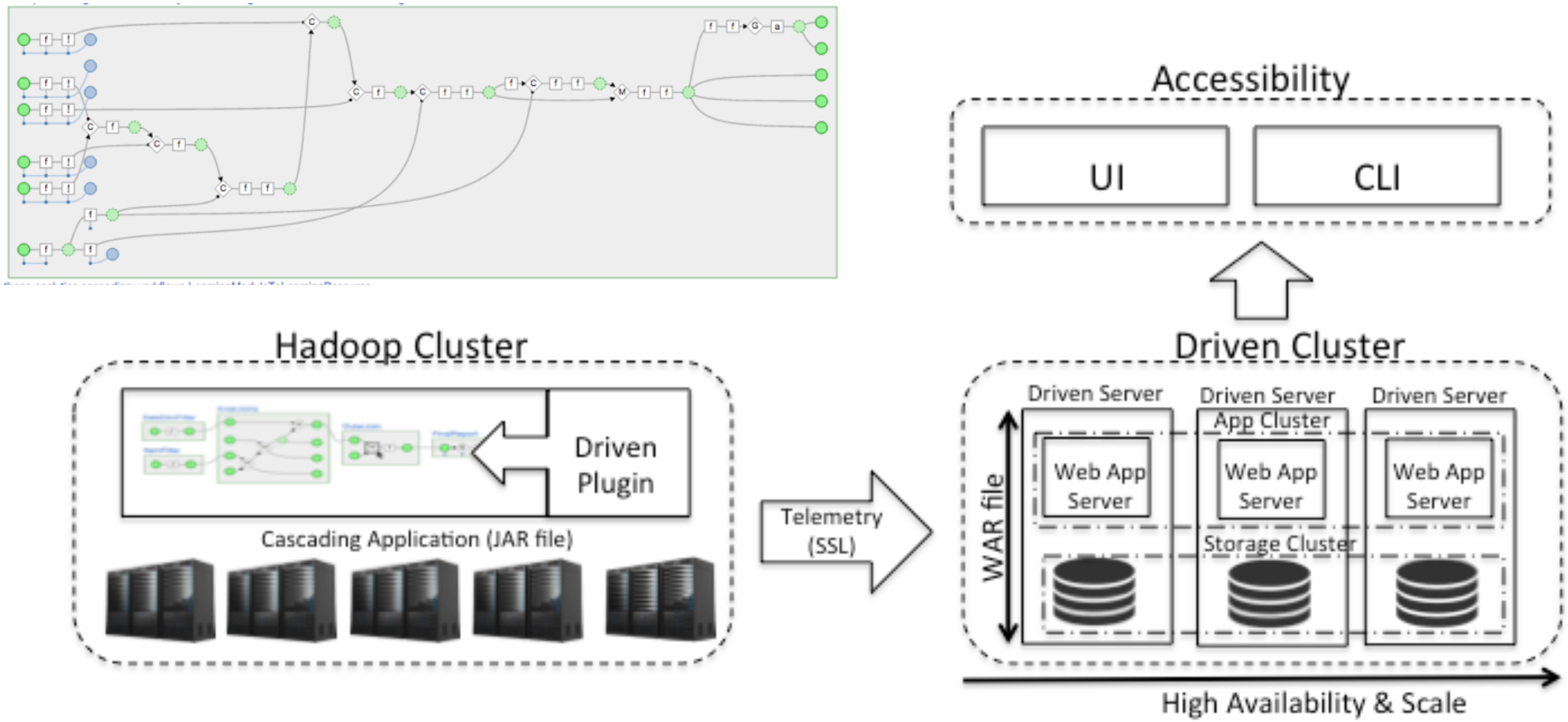
Production — Monitoring and Tracking

- Maintain Business SLAs
- Balance & Controls
- Application and Data Quality
- Operational Health
- Real-time Insights

Operational Meta-data

- Automatically Collected
- Business critical meta-data
- Scalable & searchable store
- Programmatically accessible

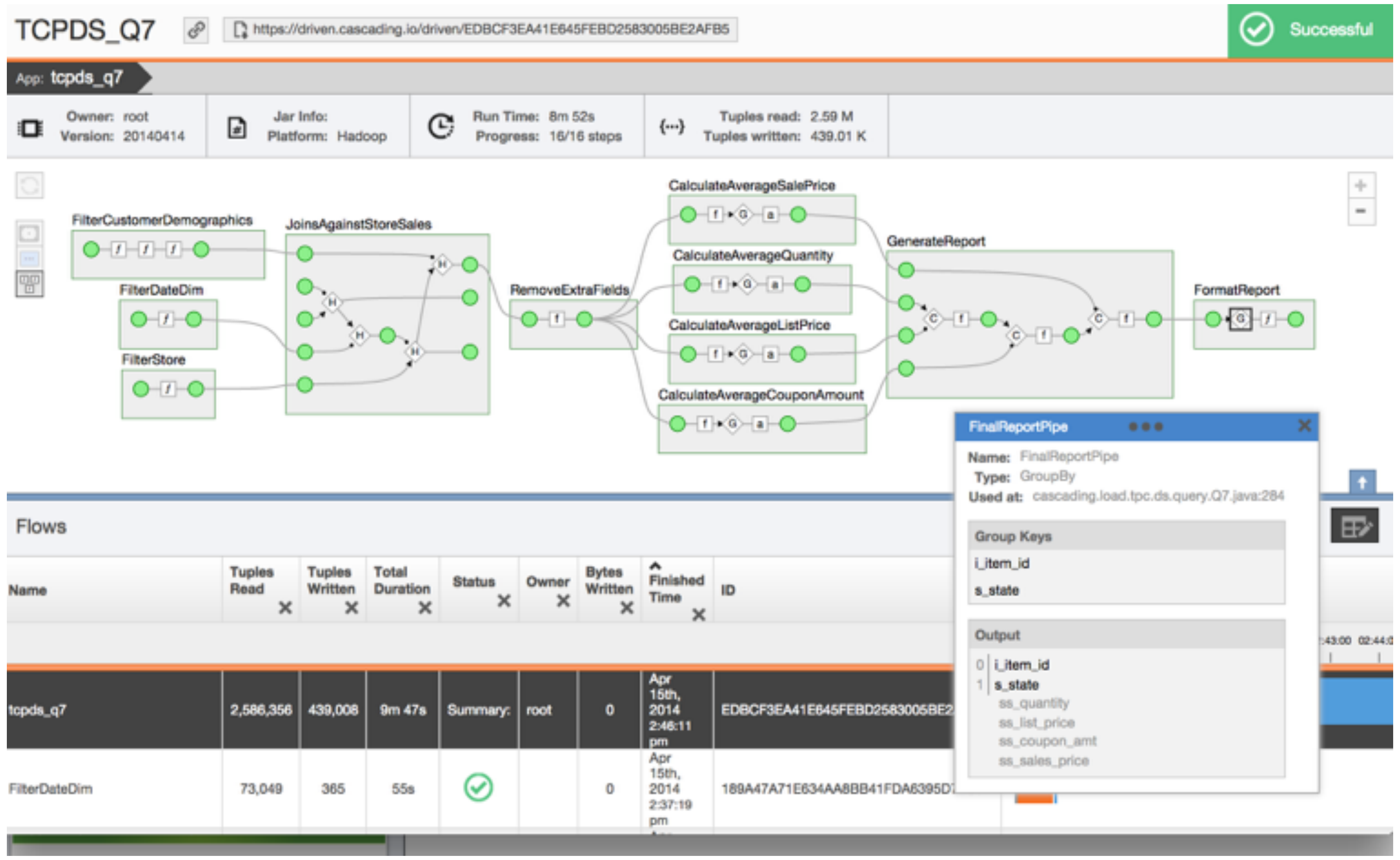
DRIVEN ARCHITECTURE



DEEPER VISUALIZATION INTO YOUR HADOOP CODE

DRIVEN

Debug and optimize your Hadoop applications more effectively with Driven



TCPDS_Q7
https://driven.cascading.io/driven/EDBCF3EA41E645FEBD2583005BE2AFB5
Successful
App: tcpds_q7
FilterCustomerDemographics
FilterDateDim
FilterStore
JoinsAgainstStoreSales
RemoveExtraFields
CalculateAverageSalePrice
CalculateAverageQuantity
CalculateAverageListPrice
CalculateAverageCouponAmount
GenerateReport
FormatReport
FinalReportPipe
Flows

Owner: root
Version: 20140414

Jar Info:
Platform: Hadoop

Run Time: 8m 52s
Progress: 16/16 steps

Tuples read: 2.59 M
Tuples written: 439.01 K

Name	Tuples Read	Tuples Written	Total Duration	Status	Owner	Bytes Written	Finished Time	ID
tcpds_q7	2,586,356	439,008	9m 47s	Summary:	root	0	Apr 15th, 2014 2:46:11 pm	EDBCF3EA41E645FEBD2583005BE2
FilterDateDim	73,049	365	55s	✓		0	Apr 15th, 2014 2:37:19 pm	189A47A71E634AA8BB41FDA6395D7

FinalReportPipe

Name: FinalReportPipe
Type: GroupBy
Used at: cascading.load.tpc.ds.query.Q7.java:284

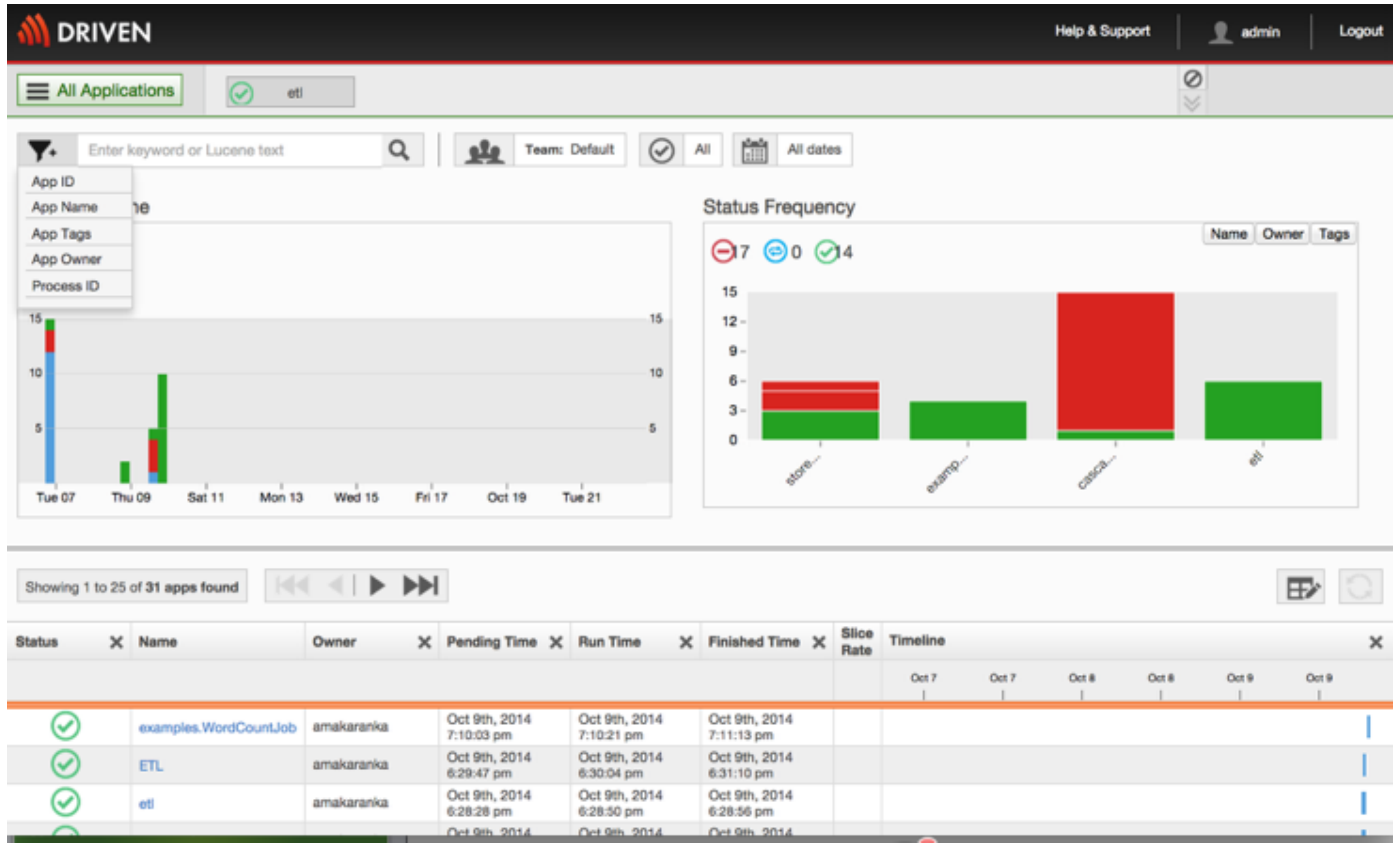
Group Keys
i_item_id
s_state

Output
0 i_item_id
1 s_state
ss_quantity
ss_list_price
ss_coupon_amt
ss_sales_price

- Easily comprehend, debug, and tune your data applications
- Get rich insights on your application performance
- Monitor *applications* in real-time
- Compare app performance with historical (previous) iterations

GET OPERATIONAL INSIGHTS WITH DRIVEN

Visualize the activity of your applications to help maintain SLAs



Showing 1 to 25 of 31 apps found

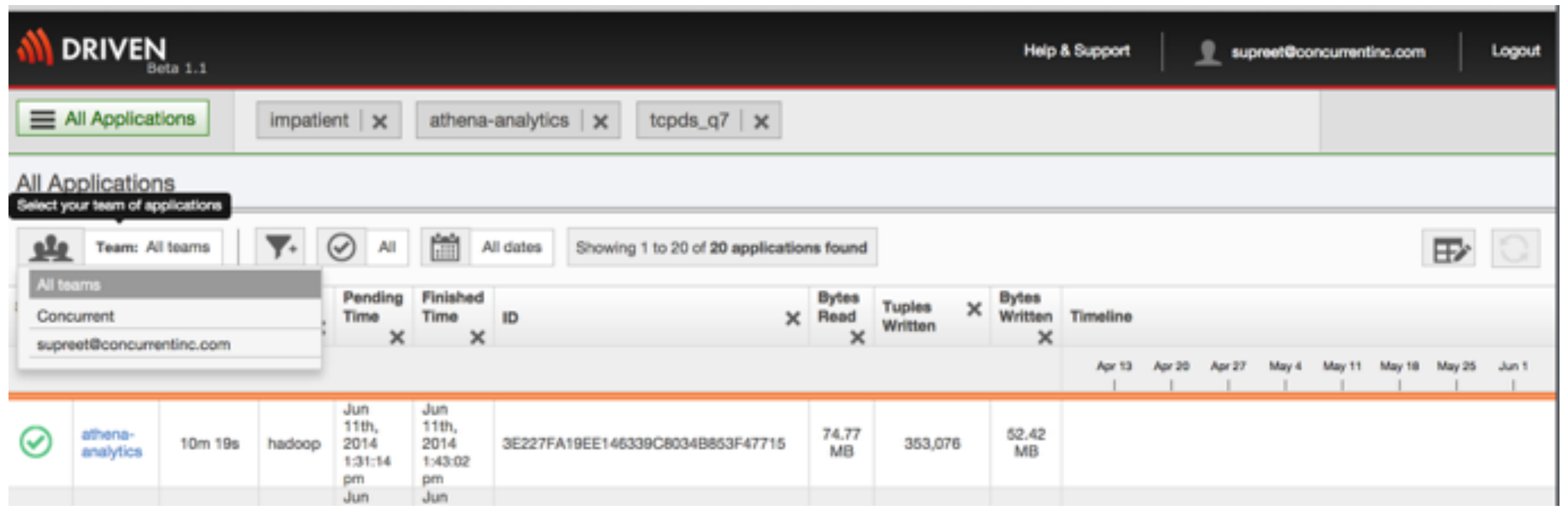
Status	Name	Owner	Pending Time	Run Time	Finished Time	Slice Rate	Timeline
✓	examples.WordCountJob	amakaranka	Oct 9th, 2014 7:10:03 pm	Oct 9th, 2014 7:10:21 pm	Oct 9th, 2014 7:11:13 pm		
✓	ETL	amakaranka	Oct 9th, 2014 6:29:47 pm	Oct 9th, 2014 6:30:04 pm	Oct 9th, 2014 6:31:10 pm		
✓	etl	amakaranka	Oct 9th, 2014 6:28:28 pm	Oct 9th, 2014 6:28:50 pm	Oct 9th, 2014 6:28:56 pm		
✓			Oct 9th, 2014	Oct 9th, 2014	Oct 9th, 2014		

- Quickly breakdown how often applications execute based on their tags, teams, or names
- Immediately identify if any application is monopolizing cluster resources
- Understand the utilization of your cluster with a timeline of all applications running

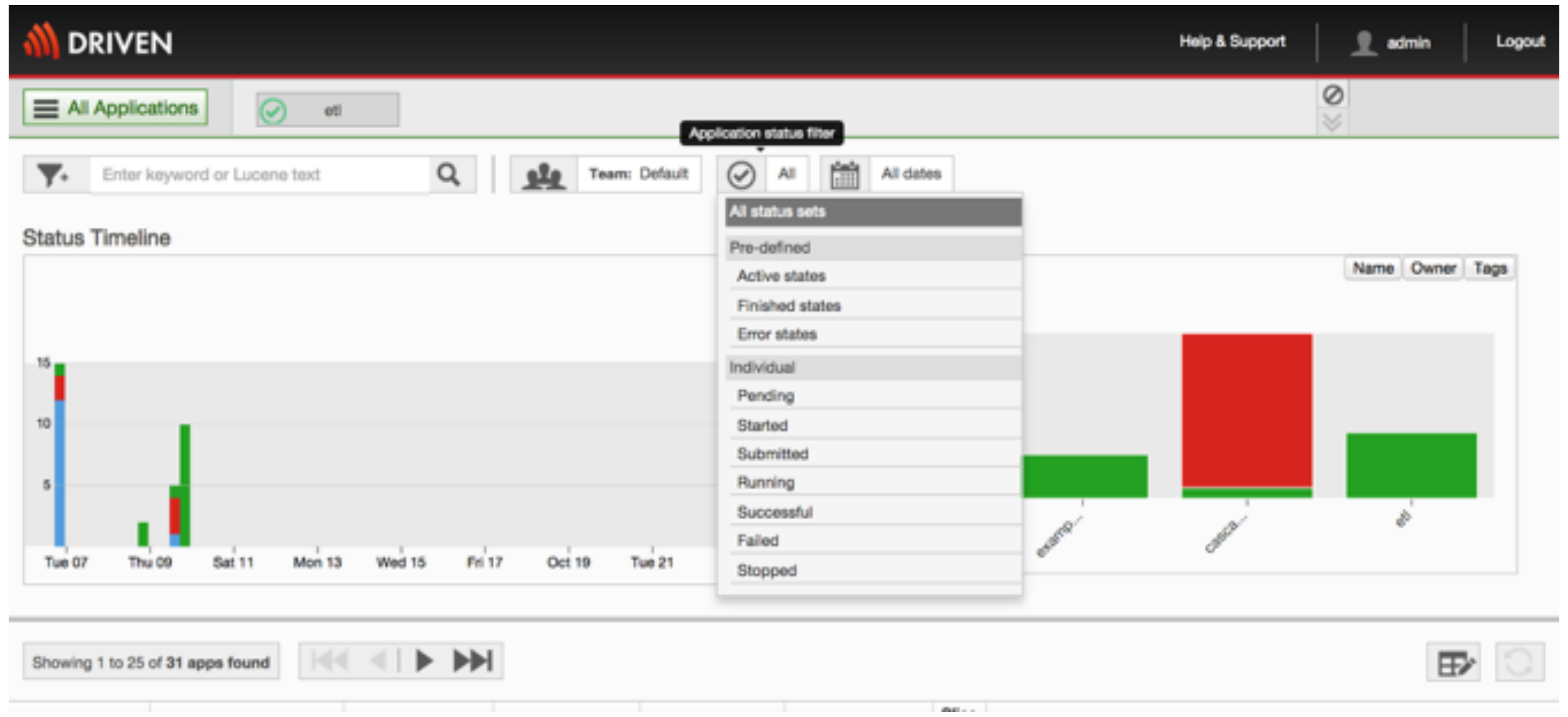
ORGANIZE YOUR APPLICATIONS WITH GREATER FIDELITY

DRIVEN

Segment your applications for greater insights across all your applications



				Pending Time	Finished Time	ID	Bytes Read	Tuples Written	Bytes Written	Timeline
✓	athena-analytics	10m 19s	hadoop	Jun 11th, 2014 1:31:14 pm	Jun 11th, 2014 1:43:02 pm	3E227FA19EE146339C8034B853F47715	74.77 MB	353,076	52.42 MB	

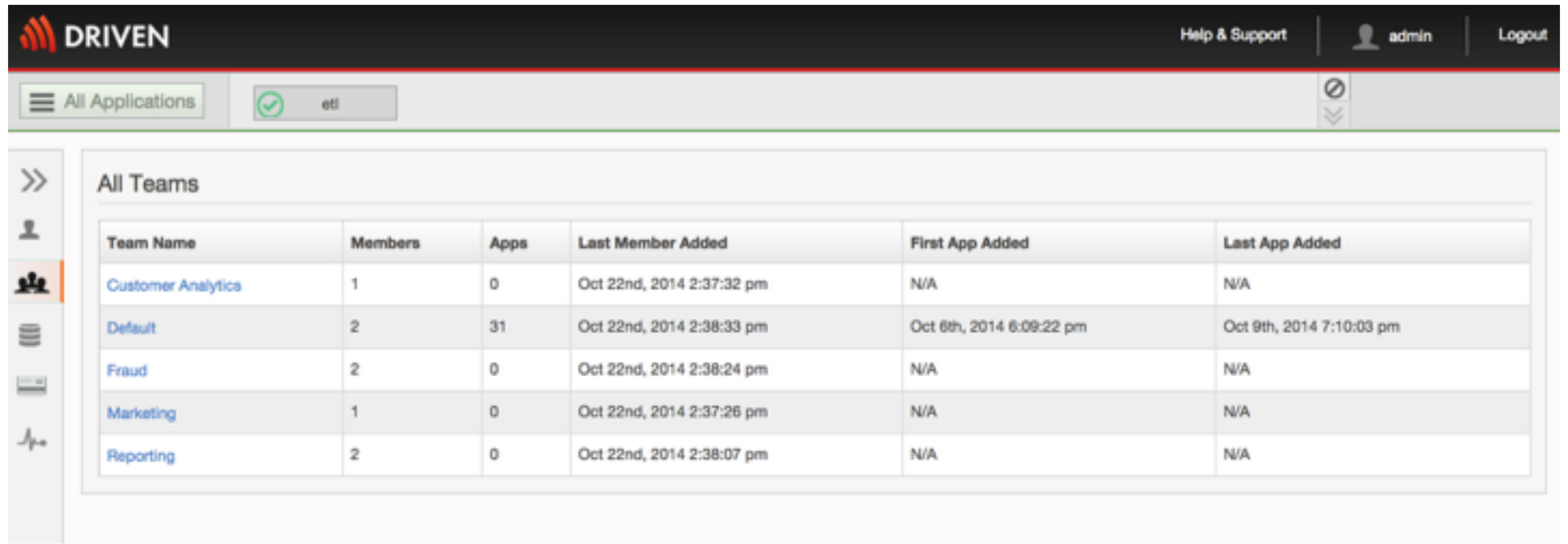


- Easily keep track of all your applications by segmenting them with user-defined tags
- Segment your applications for trending analysis, cluster analysis, and developing chargeback models
- Quickly breakdown how often applications execute based on their tags, teams, or names

CONCURRENT

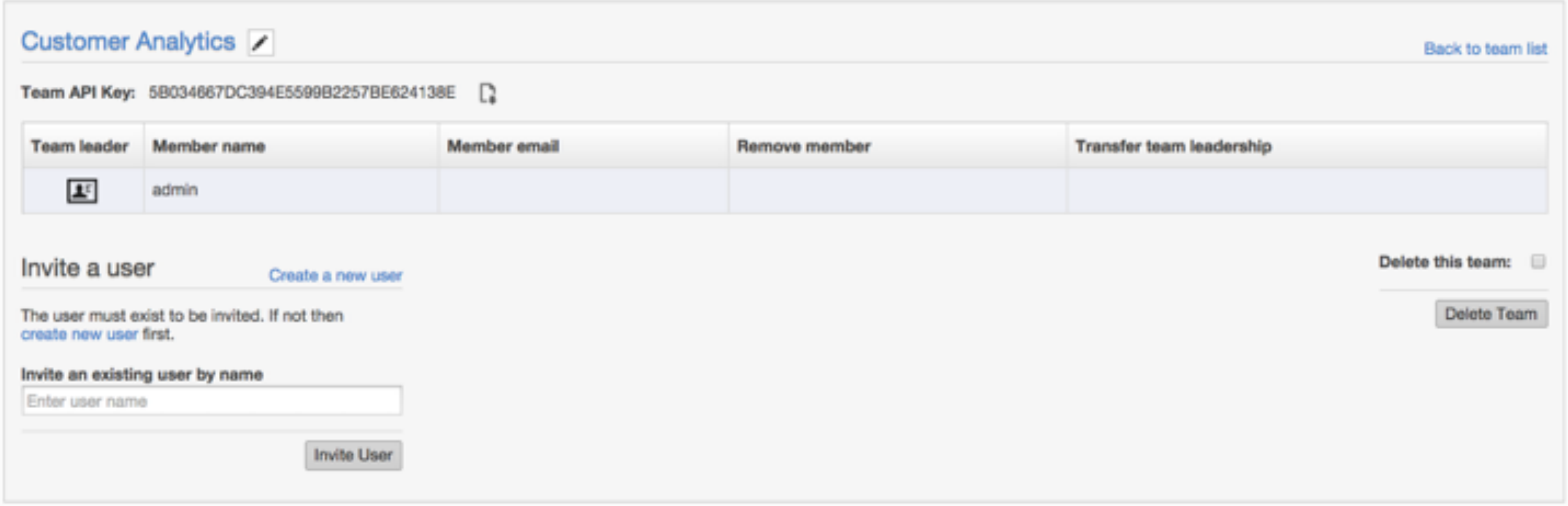
COLLABORATE WITH TEAMS

Utilize teams to collaborate and gain visibility over your set of applications



All Teams

Team Name	Members	Apps	Last Member Added	First App Added	Last App Added
Customer Analytics	1	0	Oct 22nd, 2014 2:37:32 pm	N/A	N/A
Default	2	31	Oct 22nd, 2014 2:38:33 pm	Oct 6th, 2014 6:09:22 pm	Oct 9th, 2014 7:10:03 pm
Fraud	2	0	Oct 22nd, 2014 2:38:24 pm	N/A	N/A
Marketing	1	0	Oct 22nd, 2014 2:37:26 pm	N/A	N/A
Reporting	2	0	Oct 22nd, 2014 2:38:07 pm	N/A	N/A



Customer Analytics

Team API Key: 5B034667DC394E5599B2257BE624138E

Team leader	Member name	Member email	Remove member	Transfer team leadership
👤	admin			

Invite a user

The user must exist to be invited. If not then create new user first.

Invite an existing user by name

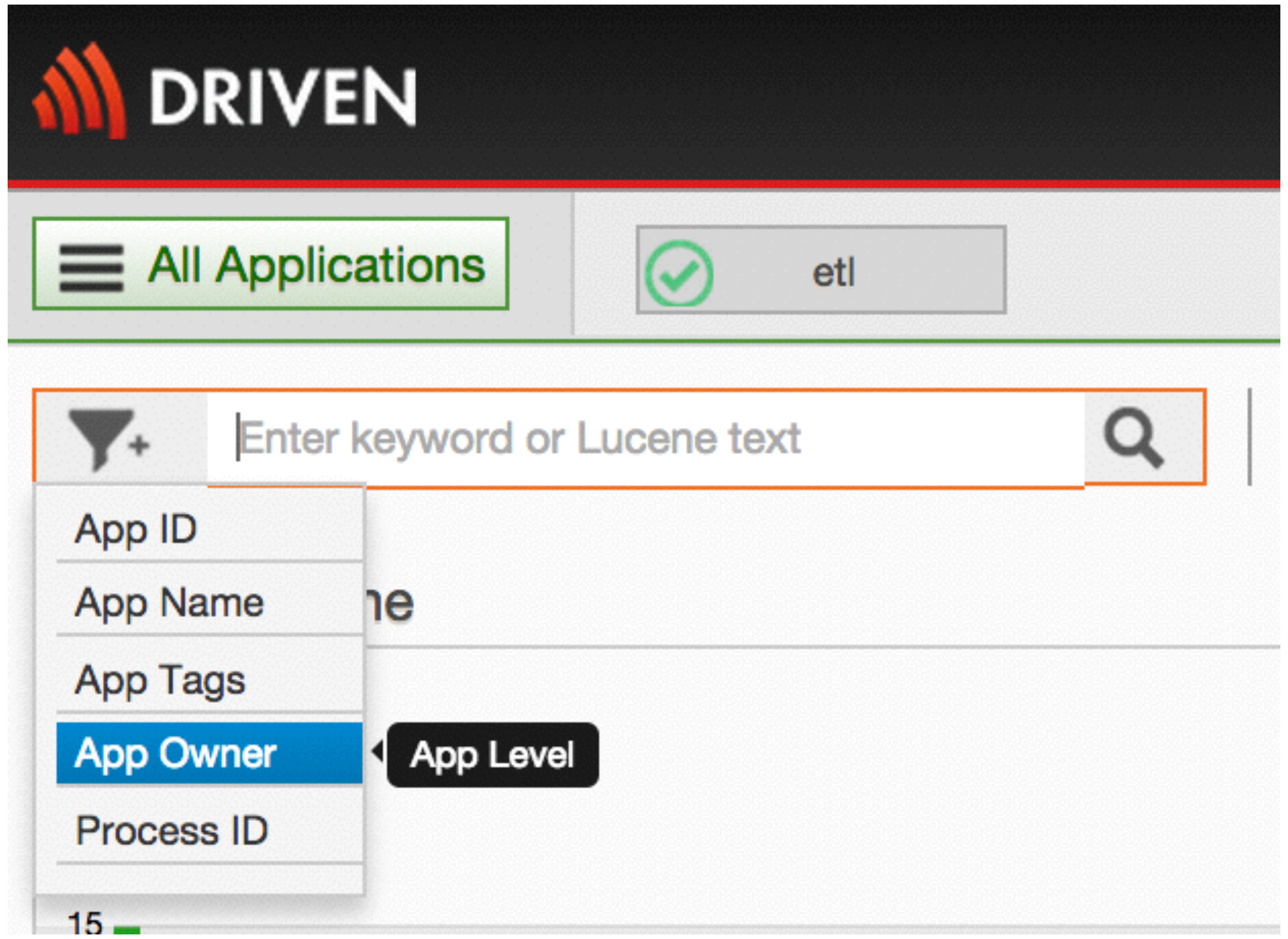
Enter user name

Delete this team: □

- Invite others to view and collaborate on a specific application
- Gain visibility to all the apps and their owners associated with each team
- Simply manage your teams and the users assigned to them

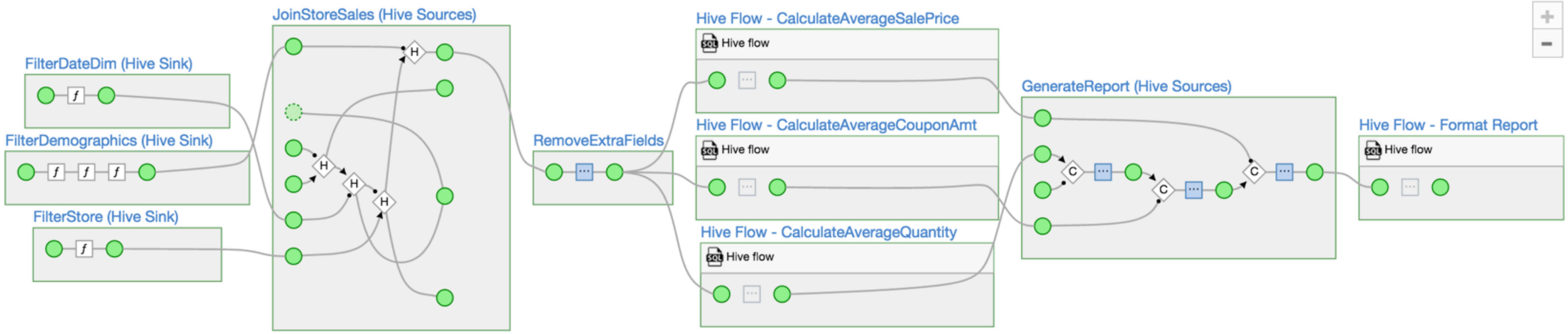
MANAGE PORTFOLIO OF BIG DATA APPLICATIONS

Fast, powerful, rich search capabilities enable you to easily find the exact set of applications that you're looking for



- Identify problematic apps with their owners and teams
- Search for groups of applications segmented by user-defined tags
- Compare specific applications with their previous iterations to ensure that your application can meet its SL

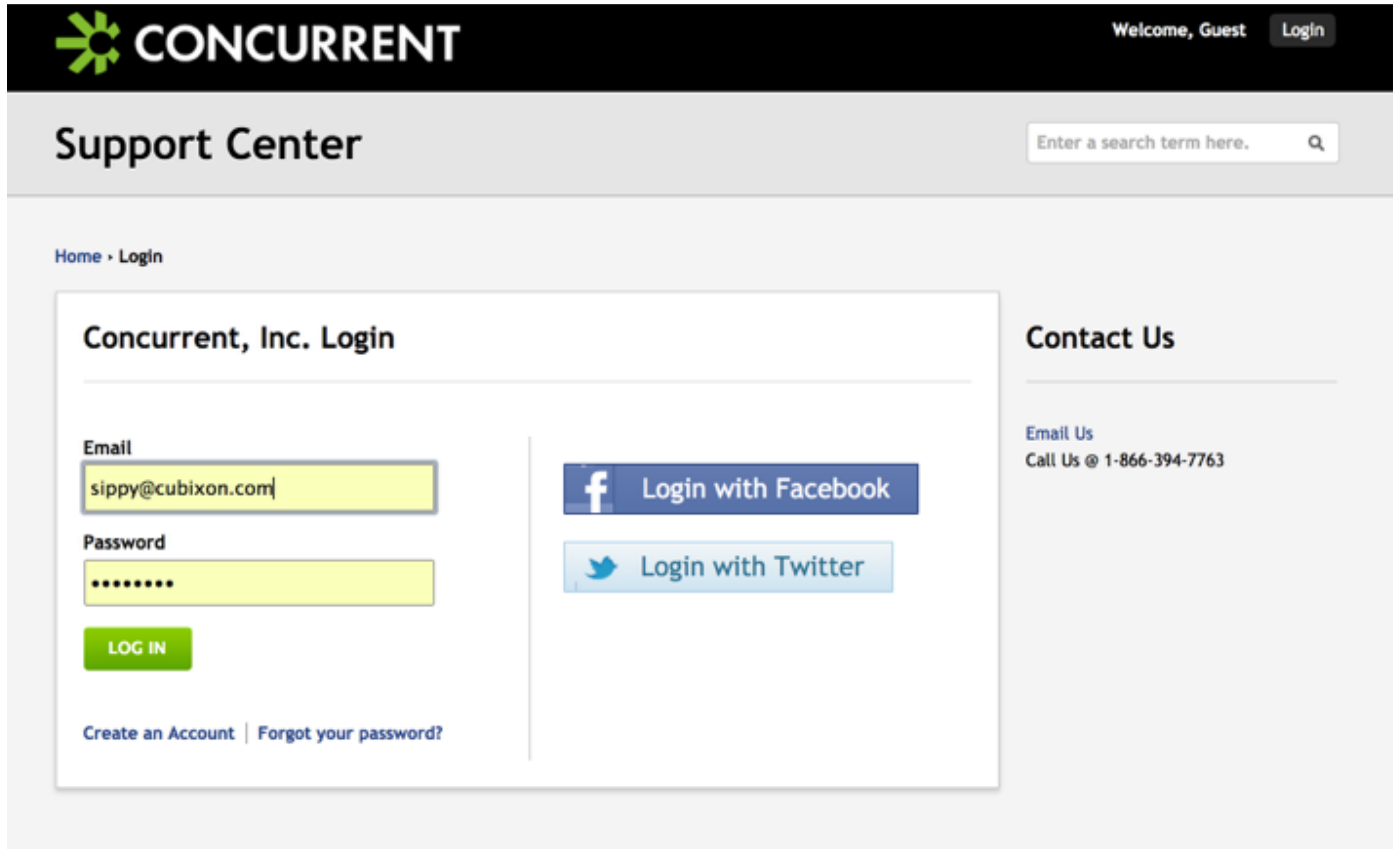
OPERATIONAL VISIBILITY FOR YOUR HIVE APPS



JoinStoreSales (Hive Sources)
FilterDateDim (Hive Sink)
FilterDemographics (Hive Sink)
FilterStore (Hive Sink)
RemoveExtraFields
Hive Flow - CalculateAverageSalePrice
Hive flow
Hive Flow - CalculateAverageCouponAmt
Hive flow
Hive Flow - CalculateAverageQuantity
Hive flow
GenerateReport (Hive Sources)
Hive Flow - Format Report
Hive flow

- Understand the anatomy of your Hive app
- Track execution of queries as single business process
- Identify outlier behavior by comparison with historical runs
- Analyze rich operational meta-data
- Correlate Hive app behavior with other events on cluster

COMMERCIAL SUPPORT FOR CASCADING



Support Center

Home › Login

Concurrent, Inc. Login

Email: sippy@cubixon.com

Password: ••••••••

Create an Account | Forgot your password?

Contact Us

Email Us

Call Us @ 1-866-394-7763

- Support for Cascading over email, phone, support portal and web forums that meet your operational SLAs
- Availability of on-site and public training classes for Cascading & Scalding
- Services of experienced technical resources provide custom design solutions
- Presence of thriving community building mission-critical applications for data-driven businesses

DRIVING INNOVATION THROUGH DATA

THANK YOU

Supreet Oberoi

CONCURRENT