DRIVING INNOVATION THROUGH DATA

USING CASCADING TO BUILD DATA-CENTRIC APPLICATIONS ON SPARK

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**
  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.

BIG DATA — OPERATIONALIZE YOUR DATA

“It’s all about the apps”

There needs to be a comprehensive solution for building, deploying, running and managing this new class of enterprise applications.

Business Strategy

Connecting Business and Data

Data & Technology

Challenges

Skill sets, systems integration,
standard op procedure and
operational visibility

DATA APPLICATIONS - ENTERPRISE NEEDS

Enterprise Data Application 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

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, SQL, modeling skill sets

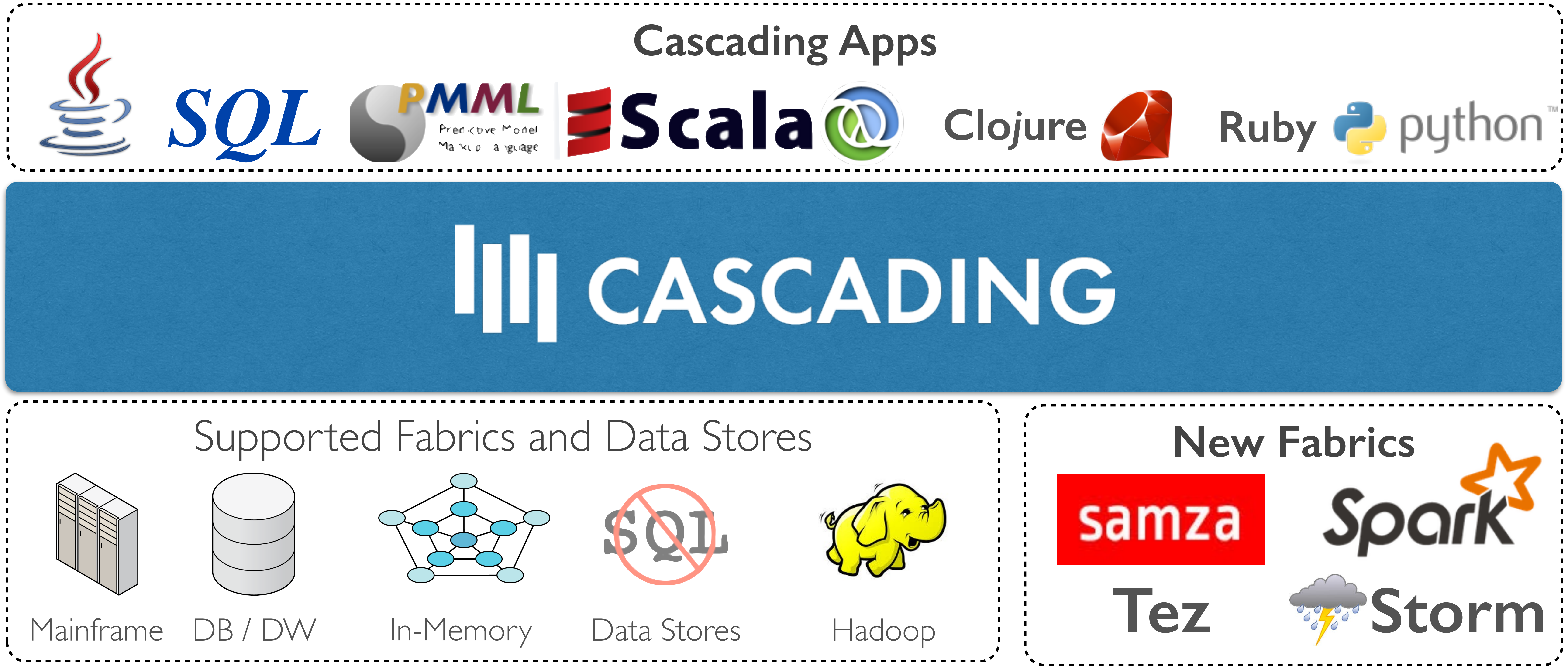
Test-Driven Development

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

Operational Complexity

Simple - Package up into one jar and hand to operations

CASCADING - DE-FACTO FOR DATA APPS



Cascading Apps

SQL PMML Predictive Model Markup Language Scala Clojure Ruby python

CASCADING

Supported Fabrics and Data Stores

Mainframe DB / DW In-Memory Data Stores Hadoop

New Fabrics

samza Spark Tez Storm

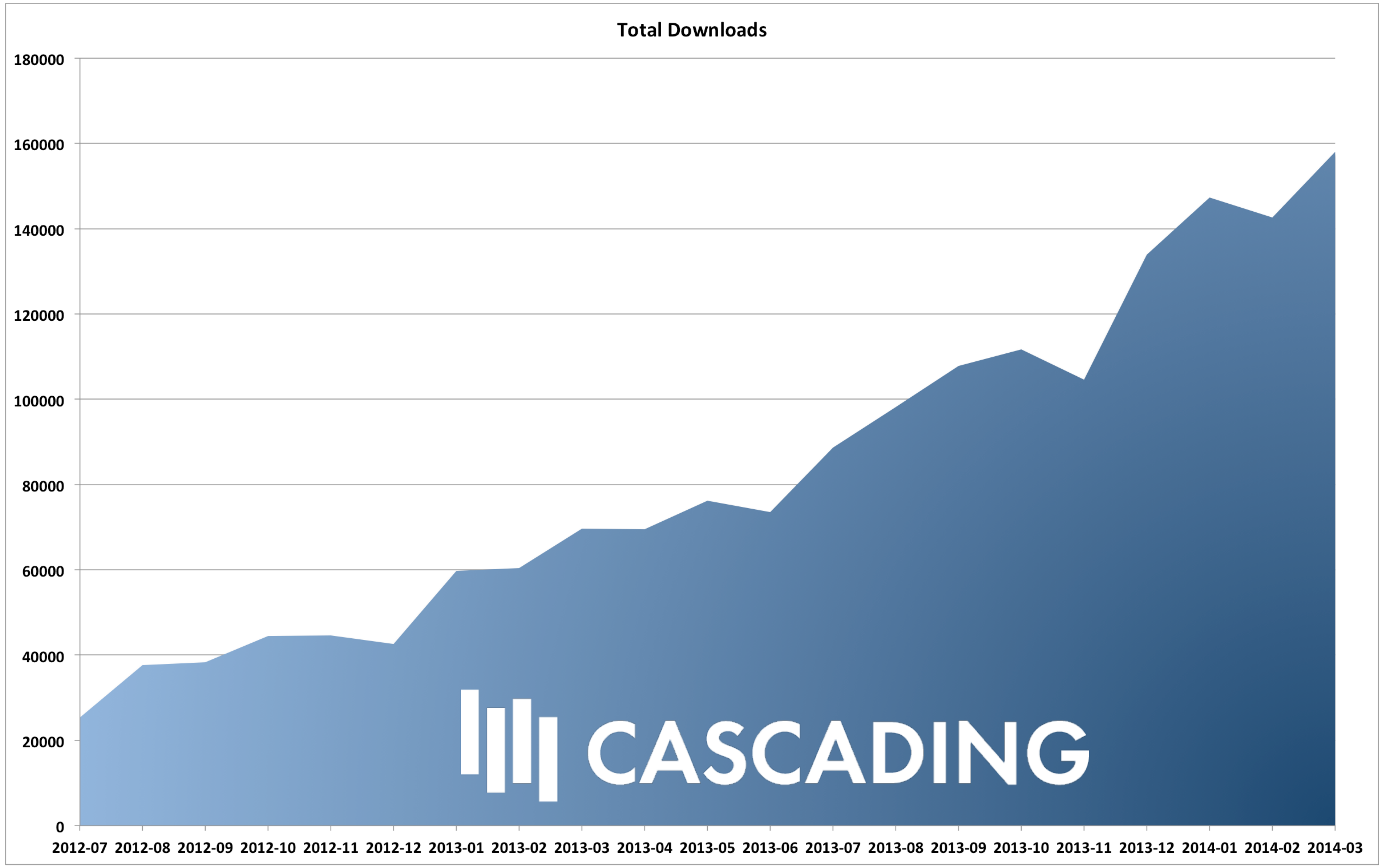
- Standard for enterprise data app development
- Your programming language of choice
- Cascading applications that run on MapReduce will also run on Apache Spark, Storm, and …

STRONG ORGANIC GROWTH

CASCADING

175,000+ downloads / month

7000+ Deployments



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

BUSINESSES DEPEND ON US

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

TWITTER

- Scalding (Scala)
- Makes complex analysis of very large data sets simple
- Machine learning, linear algebra to improve
- User experience
- Ad quality (matching users and ad effectiveness)
- All revenue applications are running on Cascading/Scalding

BUSINESSES DEPEND ON US

DURKHEIM PROJECT

- Estimate suicide risk from what people write online
- Cascading + Cassandra
- You can do more than optimize add yields
- http://www.durkheimproject.org

CASCADING DATA APPLICATIONS

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

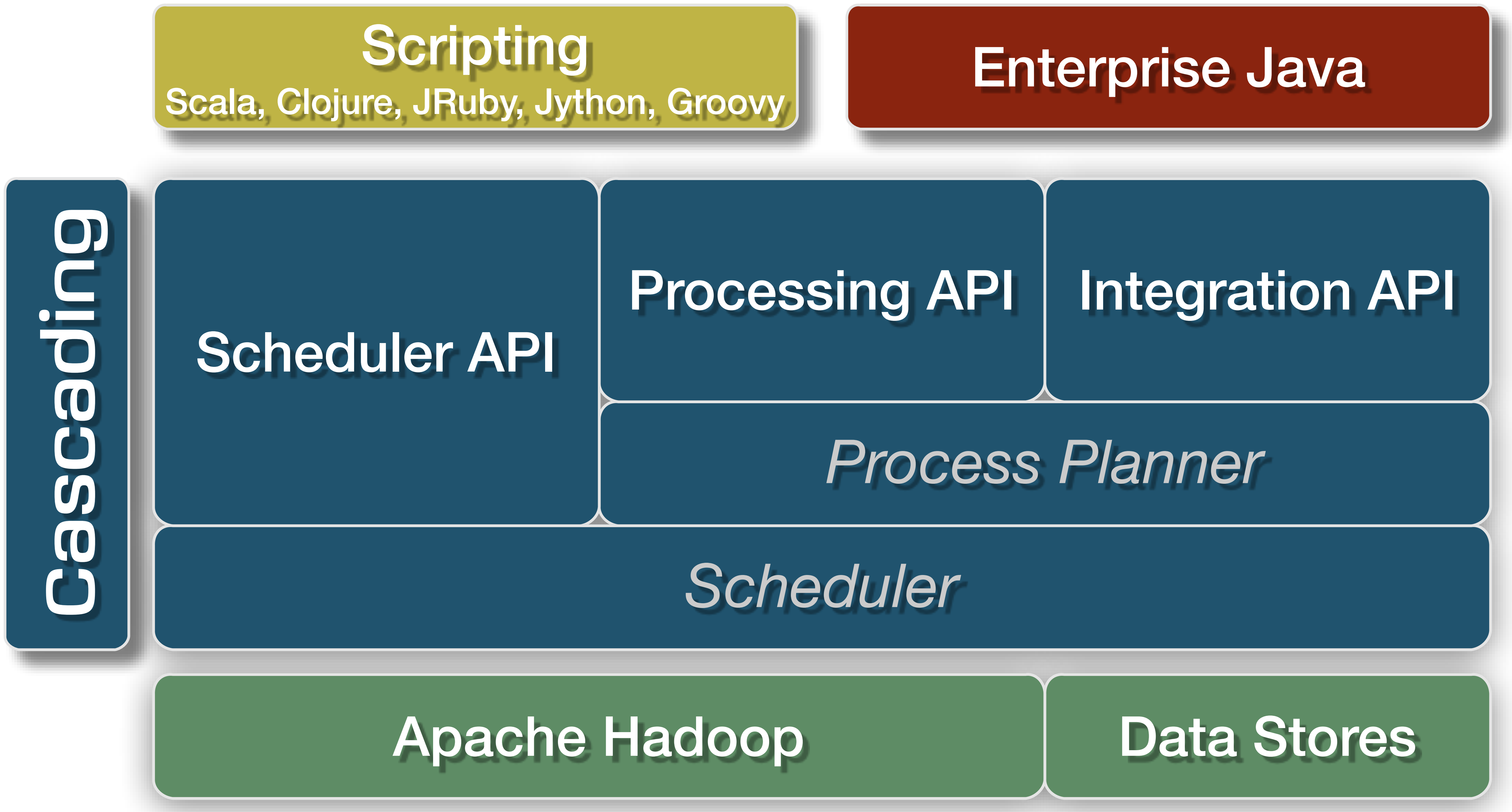
Fraud and Anomaly Detection
Fraud Experiments
Customer Analytics
Insurance Risk Metric

Health / Biotech

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

CASCADING

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



WORD COUNT EXAMPLE

```
  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 );
```

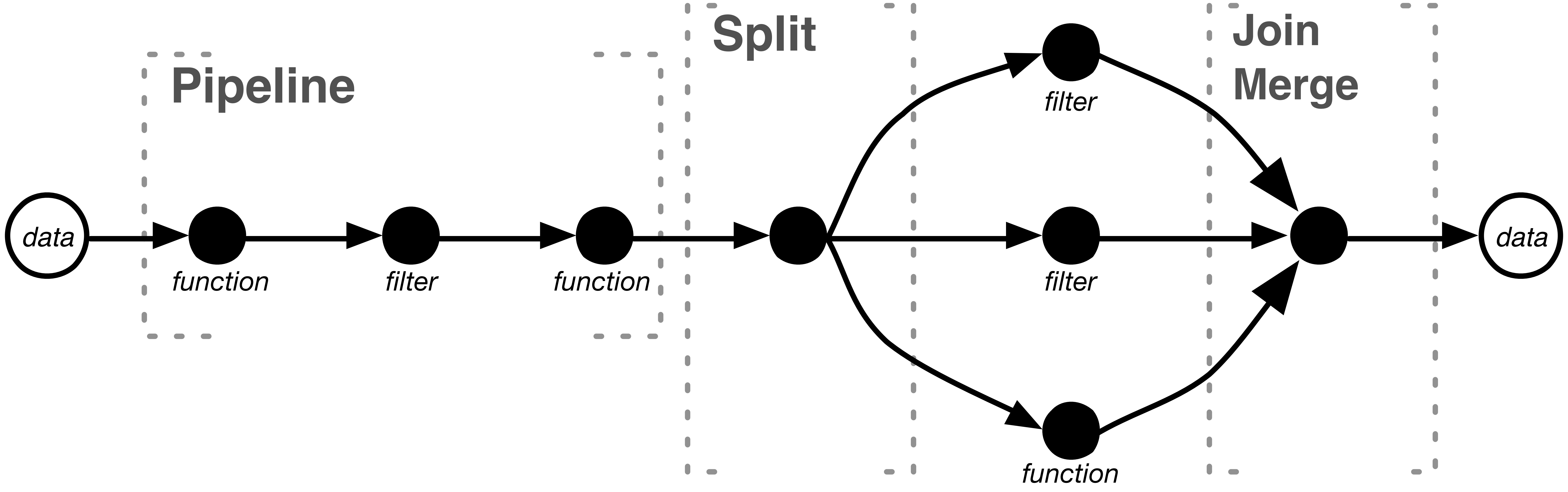
processing

```
// 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

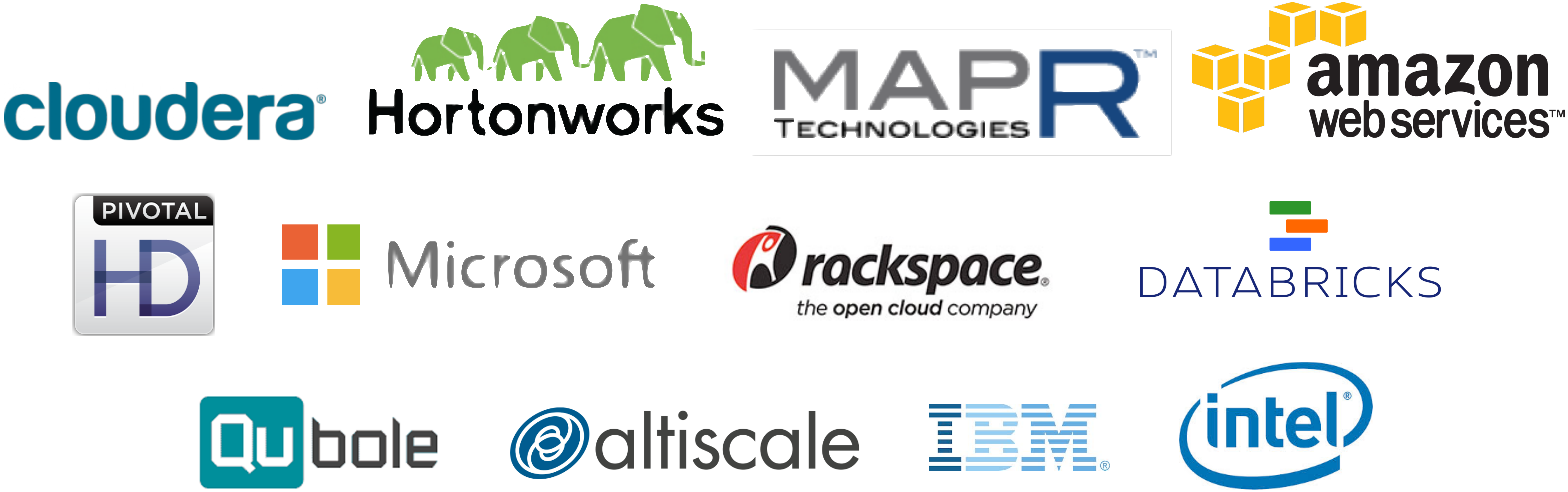
- Functions
- Filters
- Joins
  - Inner / Outer / Mixed
  - Asymmetrical / Symmetrical
- Merge (Union)
- Grouping
  - Secondary Sorting
  - Unique (Distinct)
- Aggregations
  - Count, Average, etc



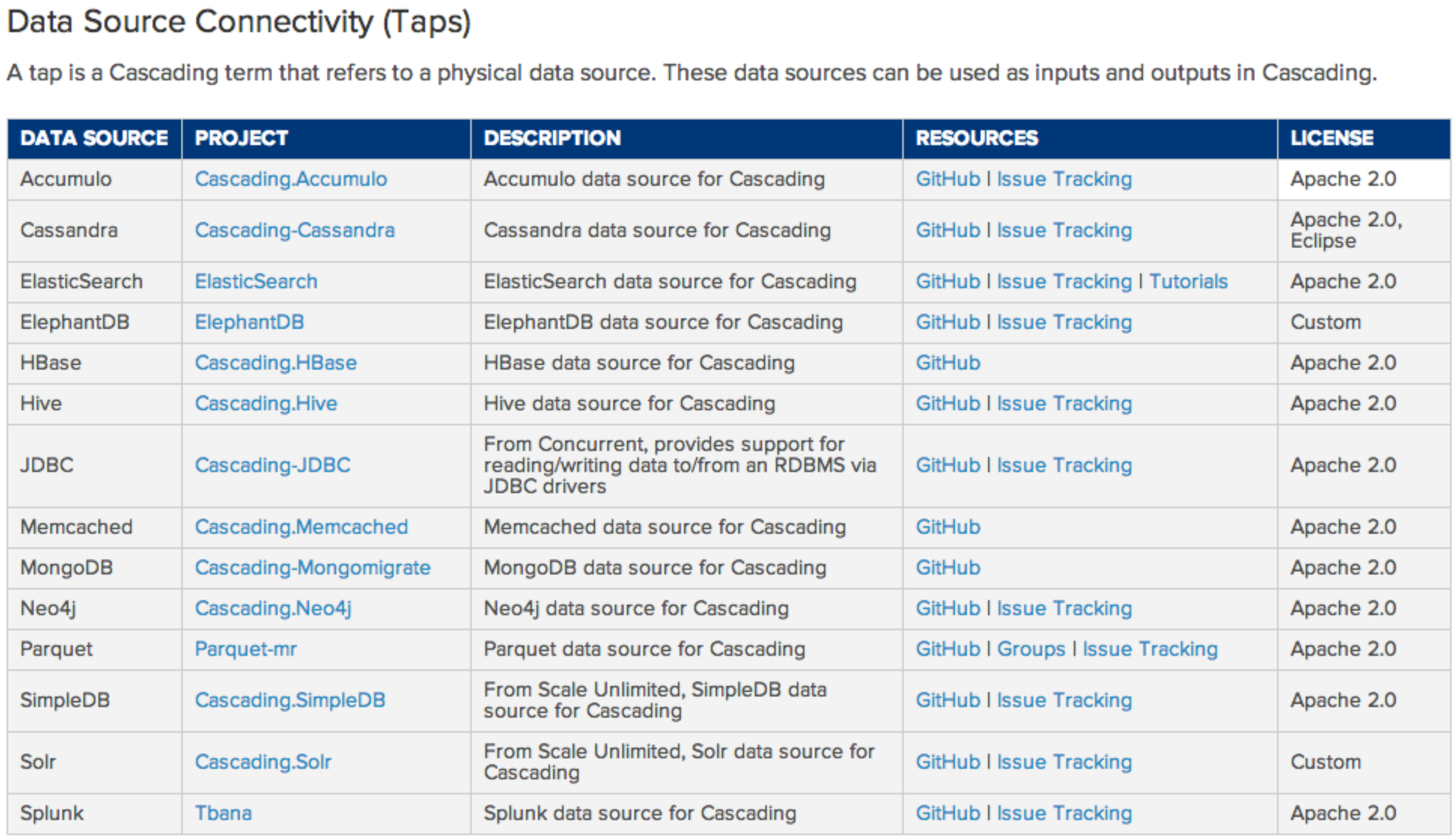
Topology

BROAD SUPPORT

Hadoop ecosystem supports Cascading



… AND INCLUDES RICH SET OF EXTENSIONS



Data Source Connectivity (Taps)

A tap is a Cascading term that refers to a physical data source. These data sources can be used as inputs and outputs in Cascading.

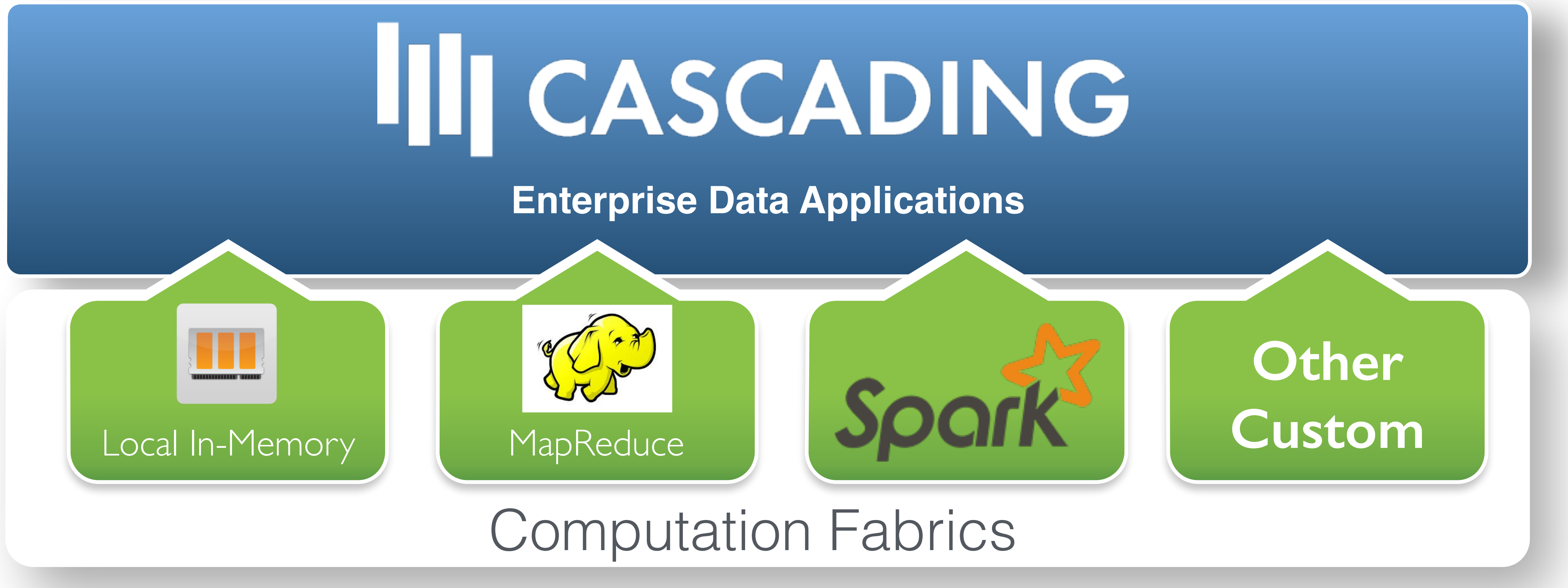
DATA SOURCE	PROJECT	DESCRIPTION	RESOURCES	LICENSE
Accumulo	Cascading.Accumulo	Accumulo data source for Cascading	GitHub I Issue Tracking	Apache 2.0
Cassandra	Cascading-Cassandra	Cassandra data source for Cascading	GitHub I Issue Tracking	Apache 2.0, Eclipse
ElasticSearch	ElasticSearch	ElasticSearch data source for Cascading	GitHub I Issue Tracking I Tutorials	Apache 2.0
ElephantDB	ElephantDB	ElephantDB data source for Cascading	GitHub I Issue Tracking	Custom
HBase	Cascading.HBase	HBase data source for Cascading	GitHub	Apache 2.0
Hive	Cascading.Hive	Hive data source for Cascading	GitHub I Issue Tracking	Apache 2.0
JDBC	Cascading-JDBC	From Concurrent, provides support for reading/writing data to/from an RDBMS via JDBC drivers	GitHub I Issue Tracking	Apache 2.0
Memcached	Cascading.Memcached	Memcached data source for Cascading	GitHub	Apache 2.0
MongoDB	Cascading-Mongomigrate	MongoDB data source for Cascading	GitHub	Apache 2.0
Neo4j	Cascading.Neo4j	Neo4j data source for Cascading	GitHub I Issue Tracking	Apache 2.0
Parquet	Parquet-mr	Parquet data source for Cascading	GitHub I Groups I Issue Tracking	Apache 2.0
SimpleDB	Cascading.SimpleDB	From Scale Unlimited, SimpleDB data source for Cascading	GitHub I Issue Tracking	Apache 2.0
Solr	Cascading.Solr	From Scale Unlimited, Solr data source for Cascading	GitHub I Issue Tracking	Custom
Splunk	Tbana	Splunk data source for Cascading	GitHub I Issue Tracking	Apache 2.0

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

CASCADING 3.0 - CURRENTLY WIP

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

- The Innovation — Cascading 3.0 will allow for data apps to execute on existing and emerging fabrics through its new customizable query planner.
- Cascading 3.0 will support — Local In-Memory, Apache MapReduce and soon thereafter (3.1) Apache Spark and Apache Storm



CASCADING 3.0 IMPACT - DATA APP DEVELOPMENT FOR SPARK ON ROBUST FRAMEWORK

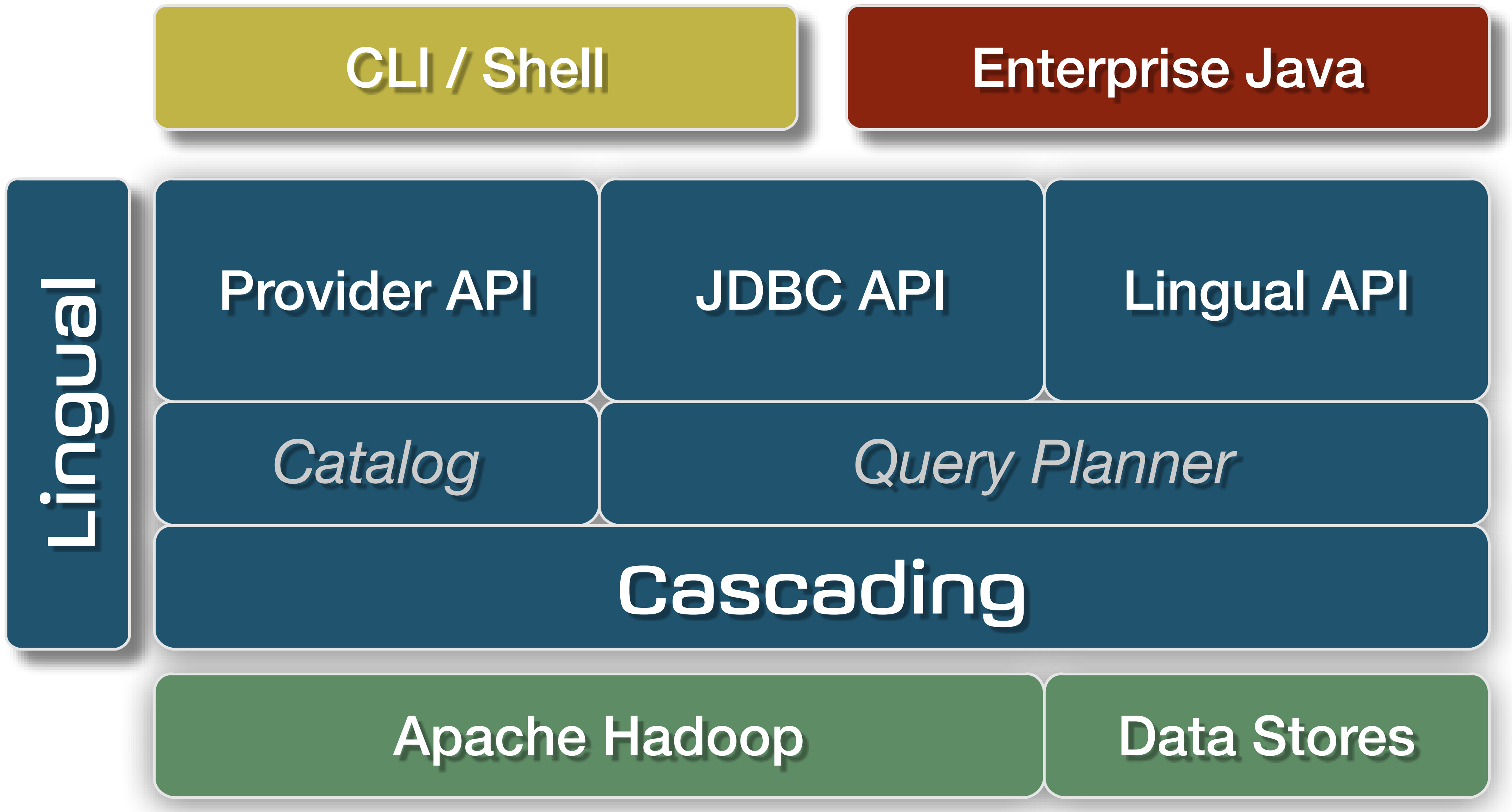
- Cascading 3.0 will ease application migration to Spark
- Enterprises can standardize on one API to meet business challenges and solve a variety of business problems ranging from simple to complex, regardless of latency or scale
- Third party products, data apps, frameworks and dynamic programming languages on Cascading will immediately benefit from this portability
- Even more operational visibility from development through production with Driven

LINGUAL

- Lingual is an extension to Cascading that executes ANSI SQL queries as Cascading apps

- Supports integrating with any data source that can be accessed through JDBC — Cascading Tap can be created for any source supporting JDBC

- Great for migration of data, integrating with non-Big Data assets — extends life of existing IT assets in an organization



SCALDING

- Scalding is a language binding to Cascading for Scala
  - The name Scalding comes from the combining of SCALa and cascaDING

- Scalding is great for Scala developers; can crisply write constructs for matrix math…

- Scalding has very large commercial deployments at:
  - Twitter - Use cases such as the revenue quality team, ad targeting and traffic quality
  - Ebay - Use cases include search analytics and other production data pipelines

PATTERN

- Pattern is an open source project that allows to leverage Predictive Model Markup Language (PMML) models and translate them into Cascading apps.
  - PMML is an XML-based popular analytics framework that allows applications to describe data mining and machine learning algorithms
- PMML models from popular analytics frameworks can be reused and deployed within Cascading workflows
  - Vendor frameworks - SAS, IBM SPSS, MicroStrategy, Oracle
  - Open source frameworks - R, Weka, KNIME, RapidMiner
- Pattern is great for migrating your model scoring to Hadoop from your decision systems

PATTERN: ALGOS IMPLEMENTED

- Hierarchical Clustering
- K-Means Clustering
- Linear Regression
- Logistic Regression
- Random Forest

algorithms extended based on customer use cases –

BUILDING AND RUNNING PMML MODELS

Training
LINGUAL
CASCADING
Data
Data
Data
ETL, prepare data
Model Producer
Explore data and build model using Regression, clustering, etc.
PMML Model
Scoring
LINGUAL
CASCADING
PATTERN
New Data
ETL, prepare data
Model Consumer
PMML model
scores
Post Processing
Measure and improve model

SPARK SUITED FOR MANY CASCADING USE CASES

- Pattern + Spark for efficiently scoring models at scale
- Lingual + Spark to efficiently cleanse and enrich data
- Cascading + Spark enables many stream processing (IoT..) and event-trigger use cases (fraud detection)
- Scalding + Spark ideal for running ML algebra & matrix math

Visibility Through All Stages of App Lifecycle

From Development — Building and Testing

- Design & Development
- Debugging
- Tuning

To Production — Monitoring and Tracking

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

SUMMARY

- Cascading framework enables developers to intuitively create data applications that scale and are robust, future-proof, supporting new execution fabrics without requiring a code rewrite
- Pattern — a Cascading extension — lets you score models at scale on Big Data fabrics, including (in near future) on Spark
- Driven — an application visualization product — provides rich insights into how your applications executes, improving developer productivity by 10x
- Cascading 3.0 opens up the query planner — write apps once, run on any fabric

Looking for Cascading-Spark contributors

CONTACT INFORMATION

Supreet Oberoi

supreet@concurrentinc.com

650-868-7675 (m)

@supreet_online

DRIVING INNOVATION THROUGH DATA

THANK YOU

Supreet Oberoi

CONCURRENT