史上最全的大数据学习资源(Awesome Big Data)

为了让大家更好地学习交流，过往记忆大数据花了一个周末的时间把 Awesome Big Data 里近600 个大数据相关的调度、存储、计算、数据库以及可视化等介绍全部翻译了一遍，供大家学习交流。

关系型数据库管理系统

- MySQL 世界上最流行的开源数据库。
- PostgreSQL 世界上最先进的开源数据库。
- Oracle Database - 对象关系数据库管理系统。
- Teradata - 高性能 MPP 数据仓库平台。

框架

- Bistro - 用于批处理和流分析的通用数据处理引擎。它基于一种新的数据模型，该模型通过函数来表示数据，并通过列操作来处理数据，而不仅仅使用 MapReduce 或 SQL 等传统方法来设置操作。
- IBM Streams - 分布式处理和实时分析平台。可以和大数据生态系统中的许多流行技术(Kafka、HDFS、Spark等) 集成
- Apache Hadoop -分布式处理框架。集成了MapReduce(并行处理)、YARN(作业调度)和HDFS(分布式文件系统)。
- Tigon - 高吞吐的实时流处理框架。
- Pachyderm - Pachyderm 是一个基于 Docker 和 Kubernetes 的数据存储平台，可以用在重复的数据处理和分析场景。
- Polyaxon - 一个可复制、可扩展的机器学习和深度学习平台。

分布式编程

- AddThis Hydra - 分布式数据处理和存储系统，最初由 AddThis 开发。
- AMPLab SIMR - 在 Hadoop MapReduce v1 上运行 Spark。
- Apache APEX - 用于大数据流和批处理的统一企业平台。
- Apache Beam - 用于定义和执行数据处理工作流的统一模型和一组特定于语言的sdk。
- Apache Crunch - 一个简单的Java API，用于处理 Join 和数据聚合之类的任务，这些任务在普通 MapReduce 上实现起来很繁琐。
- Apache DataFu - 由 LinkedIn 为 Hadoop 和 Pig 开发的用户定义函数的集合。
- Apache Flink - 分布式处理引擎框架，用于在无界和有界数据流上进行有状态计算。
- Apache Gearpump -基于 Akka 的实时大数据流引擎。
- Apache Gora - 内存数据模型和持久性框架。
- Apache Hama - BSP(Bulk Synchronous Parallel)计算框架。
- Apache MapReduce -在集群上使用并行分布式算法处理大型数据集的编程模型。
- Apache Pig - 用于表达 Hadoop 数据分析程序的高级语言。

- Apache REEF - 用来简化和统一低层大数据系统的保留性评估执行框架
- Apache S4 - 一个常规用途的、分布式的、可伸缩的、容错的、可插入式的平台，主要用于处理连续的数据流
- Apache Spark - 快速、通用的大规模数据处理引擎
- Apache Spark Streaming - 实时流处理引擎，属于 Spark 的一部分.
- Apache Storm - Twitter 开发的，可在 YARN 上进行流处理的框架。
- Apache Samza -基于 Kafka 和 YARN 的流处理的框架
- Apache Tez - 基于 YARN 的，可执行复杂 DAG (有向无环图)任务的应用程序框架。
- Apache Twill - YARN 上的抽象，减少了开发分布式应用程序的复杂性。
- Baidu Bigflow - 一个允许编写分布式计算程序的接口，它提供了许多简单、灵活、强大的 API 来轻松处理任何规模的数据。
- Cascalog - 数据处理和查询库。
- Cheetah - MapReduce 之上的高性能，用户自定义数据仓库。
- Concurrent Cascading - Hadoop 上的数据管理/分析框架。
- Damballa Parkour - 为 Clojure 开发的 MapReduce 库。
- Datasalt Pangool - 可替代 MapReduce 范式.
- DataTorrent StrAM -实时计算引擎，旨在以一种尽可能畅通的方式支持分布式、异步、实时的内存大数据计算，同时最小化开销和对性能的影响。
- Facebook Corona - Hadoop 的增强，可以消除单点故障。
- Facebook Peregrine - Map Reduce 框架.
- Facebook Scuba - 分布式内存数据存储。
- Google Dataflow - 创建数据管道来帮助我们摄取、转换和分析数据。
- Google MapReduce - map reduce 框架.
- Google MillWheel - 容错流处理框架。
- IBM Streams - 用于分布式处理和实时分析的平台。
  提供开箱即用的高级分析工具包，如地理空间，时间序列等。
- JAQL - 声明式编程语言，用于处理结构化、半结构化和非结构化数据。
- Kite - 一组库、工具、示例和文档，重点在于简化在 Hadoop 生态系统之上构建系统的过程。
- Metamarkets Druid - 用于实时分析大型数据集的框架。
- Netflix PigPen - 是 Clojure 语音的 Map-Reduce，可以编译到 Apache Pig 或者 Cascading 中
- Nokia Disco - 诺基亚开发的 MapReduce 框架。
- Onyx - 云的分布式计算。
- Pinterest Pinlater - 异步作业执行系统。
- Pydoop - 用 Python 编写，并采用 MapReduce 和 HDFS 技术对 Hadoop 进行扩展的 API。
- Ray - 用于构建和运行分布式应用程序的快速而简单的框架。
- Rackerlabs Blueflood - 多租户分布式度量处理系统
- Skale - NodeJS 上的高性能分布式数据处理框架。
- Stratosphere - 通用集群计算框架。
- Streamdrill - streamdrill
  在计算不同时间窗口上的事件流活动非常有用，并找出最活跃的时间窗口。
- streamsx.topology - 用于在 Java，Python 或 Scala 中构建 IBM Streams 应用程序的库。
- Tuktu - 易于使用的批处理和流式计算平台，可以使用 Scala，Akka 和 Play 构建！
- Twitter Heron - 由 Twitter 开发的一个实时、分布式、容错的流处理引擎，主要用于代替

Storm。
- Twitter Scalding - 用于 Map Reduce 作业的 Scala 库，基于 Cascading 构建。
- Twitter Summingbird - Summingbird 是一个类库，它允许我们编写看起来像原生 Scala 或 Java 集合转换的 MapReduce 程序，并在许多着名的分布式 MapReduce 平台上执行，包括 Storm 和 Scalding，由 Twitter 开发。
- Twitter TSAR - Twitter 开发的时间序列聚合器
- Wallaroo - 超快弹性数据处理引擎，可以使有状态、分析、流处理和事件驱动的 AI 应用程序能够快速投入生产，而无需考虑规模。它为开发人员提供了几种语言的 api 来实现他们的自定义业务逻辑。

分布式文件系统

- Ambry - 分布式对象存储，支持存储数万亿个小的不可变对象或者数十亿个大对象。
- Apache HDFS - 提供对应用程序数据的高吞吐量访问的分布式文件系统。
- Apache Kudu - Hadoop 的存储层可实现对数据的快速分析。
- BeeGFS - 之前称为 FhGFS，是一种并行分布式文件系统。
- Ceph Filesystem - 一个支持POSIX接口的文件系统
- Disco DDFS - 分布式文件系统。
- Facebook Haystack - 对象存储系统。
- Google Colossus - 分布式文件系统 (GFS2).
- Google GFS - 分布式文件系统。
- Google Megastore - 可扩展、高可用的存储。
- GridGain - GGFS, Hadoop 兼容的内存文件系统。
- Lustre file system - 高性能分布式文件系统。
- Microsoft Azure Data Lake Store - Azure 上兼容 HDFS 的存储
- Quantcast File System QFS - 开源分布式文件系统。
- Red Hat GlusterFS - 横向扩展网络附加的存储文件系统。
- Seaweed-FS -简单且高度可伸缩的分布式文件系统。
- Alluxio - 开源的基于内存的分布式存储系统。
- Tahoe-LAFS - 去中心化的云存储系统。
- Baidu File System - 分布式文件系统。

分布式索引

- Pilosa 开源的分布式位图索引，极大地加速了跨多个大规模数据集的查询。

文档数据模型

- Actian Versant - 面向对象的商业数据库管理系统。
- Crate Data - 是一个开源的大规模可扩展数据存储，它不需要任何管理。
- Facebook Apollo - Facebook 的类似于 Paxos 的 NoSQL 数据库。
- jumboDB - 基于 Hadoop 的面向文档的数据存储。
- LinkedIn Espresso - 可水平扩展的面向文档 NoSQL 数据存储。
- MarkLogic - 模式无关的企业 NoSQL 数据库技术。
- Microsoft Azure DocumentDB - NoSQL 云数据库服务，支持 MongoDB 协议

- MongoDB - 面向文档的数据库系统。
- RavenDB - 支持事务的开源文档数据库。
- RethinkDB - 支持表 join 和 group by 等查询的文档数据库。

Key Map 数据模型

注意:
业界存在一些术语混淆，存在两种不同的东西被称为“列式数据库”。这里列出的一些是围绕“键 - 映射”数据模型构建的分布式持久性数据库：所有数据都有一个(可能是组合的)键，键值对的映射与之关联。在某些系统中，多个这样的值映射可以与一个键关联，这些映射称为“列族”(值映射键称为“列”)。

另一种也称为“列式数据库”的技术，特点是它在磁盘或内存中如何存储数据。这些系统将所有行的相同列值数据存储在一起。因此，需要做更多的工作来获得给定键的所有列，但是需要更少的工作来获得给定列的所有值。

前一种在这里称为“键映射数据模型”。这些和 Key-value Data Model 存储之间的界限相当模糊。

后者更多地是关于存储格式而不是数据模型，这些数据库我们把它归到 Columnar Databases 里面去了。

你可以到 Prof. Daniel Abadi 的博文: 了解更多关于如何区分这两存储系统的讨论。.

- Apache Accumulo - 构建在 Hadoop 之上的分布式键值存储系统。
- Apache Cassandra - 受 BigTable 启发的、面向列的分布式数据存储。
- Apache HBase - 受 BigTable 启发的、面向列的分布式数据存储。
- Baidu Tera - 受 BigTable 启发的一种大型分布式表格存储系统，具有高性能、可伸缩等存储特点，最初的设计是为了管理万亿量级的超链和网页信息。
- Facebook HydraBase - 由 Facebook 开发的 HBase 演化版本。
- Google BigTable - 面向列的分布式数据存储。
- Google Cloud Datastore - 一个完全托管的无模式数据库，用于在 BigTable 上存储非关系数据。
- Hypertable - 受 BigTable 启发的、面向列的分布式数据存储。
- InfiniDB - 通过MySQL接口访问，并使用大规模并行处理来并行化查询。
- Tephra - 使 HBase 支持事务
- Twitter Manhattan - Twitter 开发的实时、多租户分布式数据库。
- ScyllaDB - 使用 C++ 编写的面向列的分布式数据存储，完全兼容 Apache Cassandra。

Key-value 数据模型

- Aerospike - 一个分布式，高可用的 K-V 类型的 NOSQL 数据库。提供类似传统数据库的ACID操作。
- Amazon DynamoDB - 分布式 key/value 存储, Dynamo 论文的实现。
- Badger - 一个快速、简单、高效和持久的键值存储，是用 Go 编写。
- Bolt - 可在 Go 语言中使用的嵌入式键值数据库.

- BTDB - .Net 中的 Key Value 数据库，包含 Object DB Layer, RPC, dynamic IL 等等。
- BuntDB - Go 语言的一个快速，可嵌入，基于内存的键/值数据库，支持自定义索引和地理空间。
- Edis - 协议兼容 Redis 的数据库，可替代 Redis。
- ElephantDB - 专门用于从 Hadoop 导出数据的分布式数据库。
- EventStore - 分布式时间序列数据库。
- GridDB - 一款高度可扩展的 NoSQL 数据库，非常适用于物联网和大数据领域，还具有高可靠性和高性能这些特性。
- HyperDex - 可扩展的下一代键值和文档存储，具有多种功能，包括一致性，容错性和高性能。
- Ignite - 分布式内存网格数据库，具有可持久化，分布式事务，分布式计算等特点，此外还支持丰富的键值存储以及SQL语法。
- LinkedIn Krati - 一个简单的持久化数据存储，具有非常低的延迟和高吞吐量。
- Linkedin Voldemort - 分布式 key/value 存储系统。
- Oracle NoSQL Database - Oracle 公司开发的分布式 key/value 存储系统。.
- Redis -一个开源(BSD许可)的，内存中的数据结构存储系统，它可以用作数据库、缓存和消息中间件。
- Riak - 去中心化的数据库存储。
- Storehaus -Twitter 开发的用于异步 key/value 存储的类库。
- SummitDB - 基于内存的 NoSQL 键/值数据库，具有磁盘持久性，并支持 Raft 一致性算法。
- Tarantool - 一个高效的 NoSQL 数据库和一个 Lua 应用服务器。
- TiKV - 一个基于 Rust 的分布式键值数据库，并受谷歌 Spanner 和 HBase 的启发。
- Tile38 - 具有空间索引和实时地理围栏的地理位置数据库。支持各种对象类型，包括纬度/经度点，边界框，XYZ切片，Geohashes和GeoJSON
- TreodeDB - key-value 存储，支持数据副本、分片以及提供原子多行写。

图数据模型

- AgensGraph - 基于 PostgreSQL 的新一代多模型图数据库。
- Apache Giraph - 一个可伸缩的分布式迭代图处理系统， 基于 Hadoop 平台，灵感来自 BSP (bulk synchronous parallel) 和 Google 的 Pregel。
- Apache Spark Bagel - Bagel 是谷歌 Pregel 图处理框架的 Spark 实现，支持基本的图形计算、组合器（combiners）和聚合器（aggregators）。目前已经被 GraphX 替代，在 Spark 2.0.0 版本已经被移除。
- ArangoDB - 多模型分布式数据库。
- DGraph - 一个可伸缩的、分布式的、低延迟的、高吞吐量的图数据库，旨在提供谷歌生产级别的规模和吞吐量，具有足够低的延迟，可以在 TB 级的结构化数据上为实时用户查询提供服务。
- EliasDB - 一个轻量级的基于图的数据库，不需要任何第三方库。
- Facebook TAO - TAO 是 facebook 广泛使用的分布式数据存储，用于存储和服务社交图。
- GCHQ Gaffer - Gaffer 是 GCHQ（英国政府通讯总部）于2015年12月14日在 GitHub 上公布的第一个开源项目，Gaffer 是个大规模图形数据库，可以方便存储大规模图的框架，节点和边界有数据统计，比如计数，直方图和草图。这些统计数据是时间窗口的节点和边界属性，可以根据时间动态更新。

- Google Cayley - 开源的图数据库。
- Google Pregel - 图处理框架。
- GraphLab PowerGraph - 包含 C++ 实现的 GraphLab API以及一组基于GraphLab API 构建的高性能机器学习和数据挖掘工具包。
- GraphX - 一个分布式图处理框架，它是基于 Spark 平台提供对图计算和图挖掘简洁易用的而丰富的接口，极大的方便了对分布式图处理的需求。
- Gremlin - 图遍历语言。
- Infovore - 一个 map/reduce 框架，用来处理大量的 RDF 数据集，注入 Freebase 和 DBpedia，基于 Hadoop 构建。
- Intel GraphBuilder - 基于 Hadoop 构造的大型图工具。
- JanusGraph - 开源分布式图形数据库，后端存储可以选择多种组件包括 Bigtable、HBase、Cassandra等，同时索引后端也可以选择很多种，包括 Elasticsearch、Solr、Lucene 等。
- MapGraph - 一个高级的 API 用于快速开发基于 GPU 的高性能图形分析应用。
- Microsoft Graph Engine - 一个基于内存的分布式大规模图数据处理引擎，能够帮助用户更方便地构建实时查询应用和高吞吐量离线分析平台。在此之前，它在学术界更广为人之的名称是 Trinity。
- Neo4j - 一个高性能的 NOSQL图数据库，完全由 Java 实现。
- OrientDB - 文档图形数据库。
- Phoebus - 大型图处理框架。
- Titan - 建立在 Cassandra 之上的分布式图数据库。
- Twitter FlockDB - 分布式图数据库。
- NodeXL - Microsoft® Excel® 2007, 2010, 2013 and 2016 免费开源的模板，可以很容易的探索网络图。

列式数据库

注意 请读一下 Key-Map Data Model 章节的说明。

- Columnar Storage - 解释什么是列式存储，以及我们什么时候需要它。
- Actian Vector - 面向列的分析数据库。
- C-Store - 面向列的 DBMS.
- ClickHouse - 一个开源的列式数据库（DBMS），主要用于在线分析处理查询（OLAP）。
- EventQL - 为大规模事件收集和分析而构建的分布式、面向列的数据库。
- MonetDB - 列式存储数据库。
- Parquet - 灵感来自于2010年 Google 发表的 Dremel 论文，是一种列式存储格式，与语言、平台无关，并且不需要和任何一种数据处理框架绑定。
- Pivotal Greenplum - 为特定目的而构建的专用分析数据仓库，它提供了一个列式存储引擎和一个传统的基于行的引擎。
- Vertica - 设计用于管理大量快速增长的数据，提供非常快的查询性能。
- SQream DB - 以色列大数据公司开发的跑在 GPU 上的大数据数据库，设计用于分析和数据仓库，使用 ANSI-92 SQL，适用于10TB到1PB的数据集。
- Google BigQuery - Google 推出的一项 Web 服务，该服务让开发者可以使用 Google 的架构来运行 SQL 语句对超级大的数据库进行操作。

- Amazon Redshift - 一个支持 SQL 查询的、快速、可扩展的列式存储数据库，它支持 PB 级的数量查询，是适用于企业级的数据仓库。
- IndexR - 一个开源的大数据存储格式，于 2017 年 1 月初正式开源，旨在通过添加索引、优化编码方式、提高 IO 效率等各种优化方式来提高计算层和存储层的数据交换效率，从而提升整体性能。
- LocustDB - 一个大规模并行且高性能的分析数据库 (analytics database)，可快速处理你的所有数据，目前处于实验性阶段。

NewSQL 数据库

- Actian Ingres - 商业支持，开源 SQL 关系数据库管理系统。
- ActorDB - 分布式的 SQL 数据库，可实现可伸缩的 K/V 存储系统。ActorDB 基于 Actor 计算模型，与传统的集中式数据库不同，ActorDB 由任意数量的被成为 actor 的独立和并发 SQL 数据库组成。
- Amazon RedShift -基于 PostgreSQL 的数据仓库服务。
- BayesDB - 一个贝叶斯数据库，内建贝叶斯查询语言 BQL，用户无需统计方面知识即可解决一些基本的科学数据问题
- Bedrock - 构建在 SQLite 之上的简单、模块化、网络化、分布式事务层。
- CitusDB - 通过分片和副本扩展 PostgreSQL。
- Cockroach - 可伸缩、地理复制、事务性数据存储。
- Comdb2 - 一个基于乐观并发控制技术的集群 RDBMS。
- Datomic - 分布式数据库旨在支持可伸缩、灵活和智能的应用程序。
- FoundationDB - 分布式数据库，受 F1 启发。
- Google F1 - 构建在 Spanner 之上的分布式 SQL 数据库。
- Google Spanner - Google的全球级的分布式数据库，具有可扩展，多版本，全球分布式、同步复制等特性。
- H-Store - 一个实验性的数据库管理系统。它专为驻线交易处理应用程序而设计。
- Haeinsa - Haeinsa 是 HBase 可线性扩展的多行，多表事务库。使用两阶段锁定和乐观并发控制来实现事务。事务的隔离级别是可序列化的。基于 Percolator 实现。
- HandlerSocket - MySQL/MariaDB 的 NoSQL 插件。
- InfiniSQL - 无限扩展的 RDBMS.
- Map-D - GPU 内存数据库，大数据分析可视化平台.
- MemSQL - 一款内存数据库，它通过将数据存在内存中，将 SQL 语句预编译为 C++ 而获得极速执行效率。
- NuoDB - 符合 SQL/ACID 的分布式数据库。
- Oracle TimesTen in-Memory Database - 基于内存的关系数据库管理系统，具有持久性和可恢复性。
- Pivotal GemFire XD - 低延迟、基于内存、分布式 SQL 数据存储。为内存表数据提供 SQL 接口，可在 HDFS 中持久存储。
- SAP HANA - 基于内存、面向列、关系数据库管理系统。
- SenseiDB - 分布式、实时、半结构化的数据库。
- Sky - 用于灵活、高性能的行为数据分析的数据库。
- SymmetricDS - 用于文件和数据库同步的开源软件。
- TiDB - 一款定位于在线事务处理/在线分析处理的融合型数据库产品，实现了一键水平伸

缩，强一致性的多副本数据安全，分布式事务，实时 OLAP 等重要特性。受 Google F1 启发。
- VoltDB - 声称是最快的内存数据库.

时间序列数据库

- Axibase Time Series Database - 基于 HBase 的时间序列数据库，内置可视化、规则引擎和 SQL 支持。
- Chronix - 一种时间序列存储器，用于存储高度压缩的时间序列，并支持快速访问数据。
- Cube - 使用 MongoDB 来存储时间序列数据。
- Heroic - 基于 Cassandra 和 Elasticsearch 的可扩展时间序列数据库。
- InfluxDB - 分布式时间序列数据库。
- IronDB - 可扩展、通用时间序列数据库。
- Kairosdb - 和 OpenTSDB 类似，但是构建在 Cassandra 之上。
- M3DB - 一个分布式时间序列数据库，可用于长期存储实时指标。
- Newts -基于 Apache Cassandra 的时间序列数据库。
- OpenTSDB -构建在 HBase 之上的分布式时间序列数据库。
- Prometheus - 时间序列数据库和服务监控系统。
- Beringei - Facebook 的内存时间序列数据库。
- TrailDB - 用于存储和查询一系列事件的有效工具。
- Druid MetaMarket 公司研发，专为海量数据集上的做高性能 OLAP (OnLine Analysis Processing)而设计的数据存储和分析系统
- Riak-TS Riak TS 是唯一专为物联网和时间序列数据优化的企业级 NoSQL 时间序列数据库。
- Akumuli 一个数值型时间序列数据库，可以存储、处理时序列数据
- Rhombus Cassandra的时间序列对象存储。
- Dalmatiner DB 快速分布式度量数据库
- Blueflood 一种用于摄取和处理时间序列数据的分布式系统。
- Timely 是一个时间序列数据库应用程序，它提供了基于 Accumulo 和 Grafana 的对时间序列数据的安全访问。
- SiriDB 具有集群功能的高扩展性、健壮性和快速的开源时间序列数据库。
- Thanos - Thanos 是一组组件，可以使用多个 Prometheus 部署创建具有无限存储容量的高可用度量系统。
- VictoriaMetrics - 与 Prometheus 兼容的快速，可扩展的开源 TSDB，包括单节点和群集版本。

类 SQL 处理系统

- Actian SQL for Hadoop - 高性能交互式 SQL，可以利用它访问 Hadoop 上的数据。
- Apache Drill - 一个低延迟的分布式海量数据交互式查询引擎，使用 ANSI SQL 兼容语法，本质上是一个分布式的 MPP 查询层。目的在于支持更广泛的数据源，数据格式，以及查询语言。受 Google的Dremel 启发。
- Apache HCatalog - Hadoop的表存储管理工具。
- Apache Hive - 基于 Hadoop

的一个数据仓库工具，可以将结构化数据文件映射为一张数据库表，并提供类 SQL 查询功能.
- Apache Calcite - 一款开源 SQL 解析工具, 可以将各种 SQL 语句解析成抽象语法术AST(Abstract Syntax Tree), 之后通过操作 AST 就可以把 SQL 中所要表达的算法与关系体现在具体代码之中。
- Apache Phoenix - 构建在 HBase 之上的关系型数据库层，可以对 HBase 中的数据进行低延迟访问。
- Aster Database - 类 SQL 分析处理。
- Cloudera Impala - 实时交互 SQL 大数据查询工具，受 Dremel 启发。
- Concurrent Lingual - Cascading 上的 SQL 查询语言。
- Datasalt Splout SQL - 针对大数据集的完整 SQL 查询引擎。
- Facebook PrestoDB -分布式 SQL 查询引擎。
- Google BigQuery - Google 推出的一项 Web 服务，该服务让开发者可以使用 Google 的架构来运行 SQL 语句对超级大的数据库进行操作，是 Dremel 的实现。
- PipelineDB - 一个开源的关系数据库，它可以在实时流数据上执行 SQL 查询，并将结果增量地存储在表中。
- Pivotal HDB - Hadoop 上的类 SQL 数据仓库系统。 /li>
- RainstorDB - 用于存储 PB 级结构化和半结构化数据量的数据库。
- Spark Catalyst - Apache Spark 的查询优化框架。
- SparkSQL - 使用 Spark 操作结构化的数据。
- Splice Machine - 兼具了 SQL 和 NoSQL 的各自优势，且能对操作型和分析型应用进行实时处理，具有 ACID 特性。
- Stinger - 由 Hortonworks 开发的一个彻底提升 Hive 效率的工具
- Tajo - Hadoop 之上的分布式数据仓库系统。
- Trafodion - 由惠普开发并开源的基于 Hadoop 平台的事务数据库引擎。提供了一个基于 Hadoop 平台的交易型 SQL 引擎，是一个擅长处理交易型负载的 Hadoop 大数据解决方案。

数据摄取

- Amazon Kinesis - 一种在 AWS 上流式处理数据的平台,让您可以轻松地加载和分析流数据, 同时还可让您根据具体需求来构建自定义流数据应用程序。
- Amazon Web Services Glue - 一项完全托管的提取、转换和加载 (ETL) 服务，让用户能够轻松准备和加载数据进行分析。
- Apache Chukwa - 数据采集系统。
- Apache Flume - 一个分布式的、可靠的、易用的系统,可以有效地将来自很多不同源系统的大量日志数据收集、汇总或者转移到一个数据中心存储。
- Apache Kafka - 分布式发布订阅消息系统。
- Apache NiFi - 一个易用、强大、可靠的数据处理与分发系统
- Apache Sqoop - 是一款开源的工具，主要用于在 Hadoop/Hive 与传统的数据库(Mysql、Oracle...)间进行数据的传递
- Cloudera Morphlines - 帮助将 ETL 的数据加载到 Solr、HBase 或 Hadoop 中的框架。
- Embulk - 开源的批量数据加载器，帮助在各种数据库、存储、文件格式和云服务之间传输数据。
- Facebook Scribe - 流日志数据聚合器。

- Fluentd - 用于收集事件和日志的工具。
- Google Photon - 地理分布式系统，用于实时连接多个连续流动的数据流，具有高可伸缩性和低延迟。
- Heka - 开源流处理系统。
- HIHO - 用于将不同数据源的数据和 Hadoop 进行连接的框架。
- Kestrel - 分布式消息队列系统。
- LinkedIn Databus - LinkedIn 开源的一个低延迟、可靠的、支持事务的、保持一致性的数据变更抓取系统。
- LinkedIn Kamikaze - 一种实用工具包，对 document lists 提供一系列的实现。
- LinkedIn White Elephant - 一个 Hadoop 日志收集器和展示器，它提供了用户角度的Hadoop集群可视化。
- Logstash - 一个开源的日志收集管理工具，可以采集来自不同数据源的数据,并对数据进行处理后输出到多种输出源。
- Netflix Suro - Netflix 开源的一款工具，它能够在数据被发送到不同的数据平台(如Hadoop、Elasticsearch)之前，收集不同应用服务器上的事件数据。
- Pinterest Secor - 实现 Kafka 日志持久性的服务
- Linkedin Gobblin -一套分布式数据集成框架，旨在简化大数据集成工作当中的各类常见任务，具体包括数据流与批量生态系统的提取、复制、组织与生命周期管理。
- Skizze - 一种概率数据结构服务和存储。
- StreamSets Data Collector - 使用一个简单的 IDE 来连续大数据摄取基础设施。
- Yahoo Pulsar - 由 Yahoo 开发并开源的一个企业级的发布订阅消息系统。
- Alooma - 实时的数据管道服务，支持将 MySQL 等数据源的数据移动到数据仓库中。

服务编程

- Akka Toolkit - 基于 Actor 模型，提供了一个用于构建可扩展的(Scalable)、弹性的(Resilient)、快速响应的(Responsive)应用程序的平台。
- Apache Avro - 数据序列化系统。
- Apache Curator - 为 Apache ZooKeeper 开发的类库。
- Apache Karaf - Apache 旗下的一个开源项目，同时也是一个基于 OSGi 的运行环境，Karaf 提供了一个轻量级的 OSGi 容器,可以用于部署各种组件,应用程序。
- Apache Thrift - Facebook 开源的跨语言的 RPC 通信框架
- Apache Zookeeper - 一个分布式应用程序协调服务。
- Google Chubby - 一个分布式锁服务，Chubby 底层一致性实现就是以 Paxos 为基础的
- Hydrosphere Mist - 一个将 Apache Spark 分析任务和机器学习模型转换为实时、批处理或反应性 web 服务的服务。
- Linkedin Norbert - 集群管理系统。
- Mara - 一个轻量级的自定义ETL框架。
- OpenMPI - 消息传递框架。
- Serf - 去中心化的服务发现和编排解决方案。
- Spotify Luigi - 用于构建批处理作业的复杂管道的 Python 包。它处理依赖项解析、工作流管理、可视化、处理故障、命令行集成等等。
- Spring XD - 用于数据摄取、实时分析、批处理和数据导出的分布式和可扩展系统。
- Twitter Elephant Bird - 用于处理 lzop 压缩数据的库。
- Twitter Finagle - JVM的异步网络堆栈。

调度

- Apache Airflow - Airbnb 开源的一个用 Python 编写的工作流管理平台。
- Apache Aurora - 长期运行服务和计划作业的 Mesos 框架。
- Apache Falcon - 数据管理框架。
- Apache Oozie - 工作流作业调度器。
- Azure Data Factory - 可大规模简化 ETL 的混合数据集成服务
- Chronos - 分布式和容错调度器。
- Linkedin Azkaban - 批处理工作流作业调度程序。
- Schedoscope - 用于 Hadoop 作业的敏捷调度 Scala DSL。
- Sparrow - 调度平台。

机器学习

- Azure ML Studio - 基于云的 R、Python 机器学习平台。
- brain - JavaScript 中的神经网络。
- Cloudera Oryx - 实时大规模机器学习。
- Concurrent Pattern - Cascading 上的机器学习框架。
- convnetjs - Javascript
  中的深入学习，可以在浏览器中训练卷积神经网络(或普通神经网络)。
- DataVec - 一个用于 Java 和 Scala
  深度学习的矢量化和数据预处理库。Deeplearning4j生态系统的一部分。
- Deeplearning4j - 美国 AI 创业公司 Skymind 开源并维护的一个基于 Java/JVM
  的深度学习框架，可使用CPU或GPU运行。
- Decider - Ruby中灵活且可扩展的机器学习。
- ENCOG - 支持多种高级算法的机器学习框架，以及支持规范化和处理数据的类。
- etcML -
  在线免费文本分析工具是由美国的斯坦福大学计算机教授开发的基于成熟的文本分析引擎
- Etsy Conjecture - Scalding 中可扩展的机器学习。
- Feast - 用于管理、发现和访问机器学习特性的特性存储库。Feast
  为模型训练和模型服务提供了一致的特征数据视图。
- GraphLab Create - Python 中的机器学习平台，包含大量 ML
  工具包、数据工程和部署工具。
- H2O - 使用 Hadoop、R 和 Python 进行统计、机器学习和数学运行时。
- Keras - 一个高层神经网络API，Keras 由纯 Python 编写而成并基 Tensorflow、Theano
  以及 CNTK 后端。受 Torch 启发。
- Lambdo 是一个工作流引擎，通过将一个分析管道(i)特征工程和机器学习(ii)模型训练和预测(iii)结合起来，通过用户定义(Python)函数实现表填充和列评估，大大简化了数据处理和分析。
- Mahout - 是 Apache Software Foundation(ASF) 旗下的一个开源项目,提供一些可扩展的机器学习领域经典算法的实现,旨在帮助开发人员更加方便快捷地创建智能应用程序。
- MLbase - 是Spark生态圈的一部分,专注于机器学习,包含三个组件:MLlib、MLI、ML Optimizer。
- MLPNeuralNet - 一个针对 iOS 和 Mac OS
  系统的快速多层感知神经网络库，可通过已训练的神经网络预测新实例。

- MOA - 实时进行大数据流挖掘和大规模机器学习。
- MonkeyLearn - 让文本挖掘变得很容易，可以从文本中提取和分类数据。
- ND4J - JVM 的矩阵库，可以认为是 Java 中的 Numpy。
- nupic - 一个实现了HTM学习算法的机器智能平台。
- PredictionIO - 面向开发人员和数据科学家的开源机器学习服务，构建在 Hadoop, Mahout 和 Cascading 之上。
- RL4J - 一个与 Deeplearning4j 集成的强化学习框架
- SAMOA - 分布式流数据机器学习框架。
- scikit-learn - 专门面向机器学习的 Python 开源框架，实现了各种成熟的算法。
- Spark MLlib - 使用 Spark 实现一些常见的机器学习算法和实用程序,包括分类、回归、聚类、协同过滤、降维以及底层优化,
- Sibyl - 谷歌大型机器学习系统.
- TensorFlow - 一个采用数据流图(data flow graphs)，用于数值计算的开源软件库。
- Theano - 蒙特利尔大学支持的以 Python 为核心的机器学习类库。
- Torch - 是一个基于 BSD License 的开源的机器学习的框架
- Velox - 服务于机器学习预测的系统。
- Vowpal Wabbit - 由微软和雅虎赞助的学习系统。
- WEKA - 一套机器学习软件。
- BidMach - CPU 和 GPU 加速库的机器学习库。

Benchmarking

- Apache Hadoop Benchmarking - 测试 Hadoop 性能的微基准测试。
- Berkeley SWIM Benchmark - 真实大数据工作负载基准。
- Intel HiBench - Hadoop 基准套件。
- PUMA Benchmarking - MapReduce 应用程序的基准测试套件。
- Yahoo Gridmix3 - 来自 Yahoo 工程师团队的 Hadoop 集群基准测试。
- Deeplearning4j Benchmarks

安全

- Apache Ranger - 是一个用在 Hadoop 平台上并提供操作、监控、管理综合数据安全的框架。
- Apache Eagle -由 eBay 公司开源的一个识别大数据平台上的安全和性能问题的开源解决方案。
- Apache Knox Gateway - Hadoop 集群中用于数据处理的 REST API 网关
- Apache Sentry - 为 Hadoop 集群中的元数据和数据存储提供集中、细粒度的访问控制。
- BDA - Hadoop 和 Spark 的漏洞检测器

系统部署

- Apache Ambari - 一个集中部署、管理、监控Hadoop 分布式集群的工具。
- Apache Bigtop - 一个针对基础设施工程师和数据科学家的开源项目，旨在全面打包、测试和配置领先的开源大数据组件/项目，包括但不限于 Hadoop、HBase 和 Spark 。
- Apache Helix - 集群管理框架。

- Apache Mesos - 一个类似于 YARN 的集群管理器，提供了有效的、跨分布式应用或框架的资源隔离和共享，可以运行 Hadoop、MPI、Hypertable、Spark。
- Apache Slider - 是一个 YARN 应用程序，用于在 YARN 上部署现有的分布式应用程序。
- Apache Whirr - 运行云服务的一组 Java 类库。
- Apache YARN - 集群管理系统。
- Brooklyn - 简化应用程序部署和管理的库。
- Buildoop - 类似于 Apache BigTop，基于 Groovy 语言开发。
- Cloudera HUE - 用于与 Hadoop 交互的 web 应用程序。
- Facebook Prism - 多数据中心复制系统。
- Google Borg - Google 的内部大型集群管理系统。
- Google Omega - Google 内部第三代的集群管理框架。
- Hortonworks HOYA - 可以在 YARN 上部署 HBase 集群的应用程序。
- Kubernetes - Google 团队发起并维护的基于 Docker 的开源容器集群管理系统。
- Marathon - 一个 Mesos 框架，能够支持运行长服务。

应用程序

- 411 - 一个警报管理Web应用程序。
- Adobe spindle - 使用 Scala、Spark 和 Parquet 进行 web 分析的下一代系统。
- Apache Kiji - 基于 HBase 的实时数据采集与分析框架。
- Apache Metron - 一种多功能的安全遥测数据捕获、流分析和威胁响应平台，代表了安全数据平台的最新发展水平。
- Apache Nutch - 开源 web 爬虫程序。
- Apache OODT - NASA 开源的用于做数据管理的系统。
- Apache Tika - 使用 Java 编写的内容检测和分析框架。
- Argus - 时序监控报警平台。
- AthenaX - 一个流分析平台，允许用户使用结构化查询语言(SQL)运行生产质量的大规模流分析。
- Atlas - 用于管理维度时间序列数据的系统。
- Countly - 基于 Node.js 和 MongoDB 的开源移动和 web 分析平台。
- Domino - 运行、扩展、共享和部署模型——不需要任何基础设施。
- Eclipse BIRT - 基于 Eclipse 的报告系统。
- ElastAert - 为 ES 打造的报警监控工具。
- Eventhub - 开源事件分析平台。
- Hermes - 构建在 Kafka 之上的异步消息代理。
- HIPI Library - 使用 Hadoop 的 MapReduce 来执行图像处理任务的API。
- Hunk - Hadoop 的分析工具。
- Imhotep - 大型分析平台。
- Jupyter - 基于网页的用于交互计算的应用程序。其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。
- MADlib - RDBMS 的数据处理库，用于分析数据。
- Kapacitor - 用于对时间序列数据进行处理、监视和警报的开源框架。
- Kylin - 一个开源的分布式分析引擎，提供 Hadoop/Spark 之上的 SQL 查询接口及多维分析（OLAP）能力以支持超大规模数据，最初由 eBay Inc.

开发并贡献至开源社区，能在亚秒内查询巨大的Hive表。
- PivotalR - 支持在 Pivotal HD / HAWQ 以及 PostgreSQL 上运行 R。
- Rakam - 开源实时自定义分析平台，由 Postgresql, Kinesis 和 PrestoDB 提供支持。
- Qubole - 能够自动扩展 Hadoop 集群以及内置的链接器。
- Sense - 数据科学和大数据分析的云平台。
- SnappyData - 一个统一 OLTP+OLAP +流式写入的内存分布式数据库。
- Snowplow - 由 Hadoop，Kinesis，Redshift 和 Postgres 支持的企业级 Web 和事件分析。
- SparkR - 用于 Spark 的 R 前端。
- Splunk - 一款成熟的商业化日志处理分析产品。
- Sumo Logic - 基于云的日志处理分析产品。
- Talend - YARN、Hadoop、HBASE、Hive、HCatalog 和 Pig 的统一开源环境。
- Warp - 大数据示例查询工具(OS X 应用)

搜索引擎和框架

- Apache Lucene - 一套用于全文检索和搜索的开放源码程序库
- Apache Solr - 是 Apache Lucene 项目的开源企业搜索平台。其主要功能包括全文检索、命中标示、分面搜索、动态聚类、数据库集成，以及富文本（如Word、PDF）的处理。
- Elassandra - 是 ElasticSearch 的一个分支，经过修改，可以作为 Apache Cassandra 的插件运行，具有可扩展和灵活的点对点架构。
- ElasticSearch - 一个基于 Lucene 库的搜索引擎。它提供了一个分布式、支持多租户的全文搜索引擎，具有 HTTP Web 接口和无模式 JSON 文档。
- Enigma.io – 免费增值的 Web 应用程序，用于对 Web 上抓取的海量数据集进行浏览，过滤，分析，搜索和导出。
- Facebook Unicorn - 社交图搜索平台.
- Google Caffeine - 一个高性能、出色的缓存类库。
- Google Percolator - 由 Google 公司开发的、为大数据集群进行增量处理更新的系统，主要用于 google 网页搜索索引服务。
- TeraGoogle - 大型搜索索引。
- HBase Coprocessor - HBase 的协处理器，Percolator 的实现。
- Lily HBase Indexer - 一款快速、简单的 HBase 的内容检索方案，它可以帮助你在 Solr 中建立 HBase 的数据索引，从而通过 Solr 进行数据检索。
- LinkedIn Bobo - 完全用 Java 编写的 Faceted Search 实现，是 Apache Lucene 的扩展。
- LinkedIn Cleo -一个灵活的软件库，用于处理一些预输入和自动完成的搜索功能。
- LinkedIn Galene - LinkedIn 的搜索架构。
- LinkedIn Zoie - 一个用 Java 编写的实时搜索/索引系统。
- MG4J - MG4J (Managing Gigabytes for Java) 是一个用 Java 编写的大型文档集合的全文搜索引擎，它是高度可定制的，高性能的，并提供了最先进的功能和新的研究算法。
- Sphinx Search Server - 全文搜索引擎。
- Vespa - 在大型数据集上进行低延迟计算的引擎。它存储和索引数据，以便可以在服务时执行对数据的查询，选择和处理。

MySQL 分支和演进

- Amazon RDS - AWS 的 MySQL 数据库。
- Drizzle - MySQL 6.0的演进。
- Google Cloud SQL - Google 云中的 MySQL 数据库。
- MariaDB - MySQL 的一个分支，采用GPL授权许可。目的是完全兼容 MySQL，包括 API 和命令行。
- MySQL Cluster - 使用 NDB 集群存储引擎实现 MySQL 集群。
- Percona Server - MySQL 增强版，可以替代它。
- ProxySQL - MySQL 的高性能代理。
- TokuDB - TokuDB 是 MySQL 和 MariaDB 的存储引擎。
- WebScaleSQL - WebScaleSQL 是 Facebook、 Google、Twitter 和 Linkedin 四家公司的MySQL 团队发起的 MySQL 开源组织，旨在改进 MySQL 在规模和性能等方面的问题。

PostgreSQL 分支和演进

- HadoopDB - MapReduce 和 DBMS 的混合体。
- IBM Netezza - 高性能数据仓库设备。
- Postgres-XL - 可伸缩的基于 PostgreSQL 的开源数据库集群。
- RecDB - 完全在 PostgreSQL 内部构建的开源推荐引擎。
- Stado - 仅针对数据仓库和数据集市应用程序的开源 MPP 数据库系统。
- Yahoo Everest - 由 PostgreSQL 派生的 PB 级数据库/MPP。
- TimescaleDB - 针对快速摄取和复杂查询而优化的开源时间序列数据库。
- PipelineDB - 开源的流式数据库，基于 PostgreSQL 数据库改造的，允许我们通过 SQL 的方式，对数据流做操作，并把操作结果储存起来。

Memcached 分支和演进

- Facebook McDipper - 用于闪存的键/值缓存，设计目的在于提高闪存存储的使用效率。
- Facebook Memcached - Memcache 的分支。
- Twemproxy - 一个快速、轻量级的 memcached 和 redis 代理。
- Twitter Fatcache - 用于闪存的键/值缓存。
- Twitter Twemcache - Memcache 的分支。

嵌入式数据库

- Actian PSQL - 由 Pervasive Software 开发的符合 ACID 的 DBMS，针对嵌入应用程序进行了优化。
- BerkeleyDB - 可为键/值数据提供高性能的嵌入式数据库。
- HanoiDB - Erlang LSM BTree 存储。
- LevelDB - Google 开源的持久化KV单机数据库，具有很高的随机写，顺序读/写性能。
- LMDB - 由 Symas 开发的基于 Btree-based 的高性能 mmap key-value 数据库
- RocksDB - Facebook 公司基于 LevelDB 开发的一款开源嵌入式数据库引擎。

商业智能

- BIME Analytics - 商业智能云平台。
- Blazer - 使商业智能变得简单。
- Chartio - 商业智能平台，可以可视化和浏览我们的数据。
- datapine - 自助式商业智能工具。
- GoodData - 商业智能和大数据分析软件。
- Jaspersoft - 强大的商业智能套件。
- Jedox Palo - 可定制的商业智能平台.
- Jethrodata - 交互式大数据分析。
- Metabase - 一个简单、开源的方式，通过给公司成员提问，从得到的数据中进行分析、学习。
- Microsoft - 商业智能软件及平台。
- Microstrategy - 用于商业智能、移动智能和网络应用程序的软件平台。
- Numeracy - SQL 客户端和商业智能。
- Pentaho - 商业智能平台。
- Qlik - 商业智能及分析平台。
- Redash - 开源商业智能平台，支持多个数据源和计划查询。
- Saiku - 开源分析平台。
- SpagoBI - 开源商业智能平台。
- SparklineData SNAP - 基于 Apache Spark 的商业智能平台。
- Tableau - 商业智能平台。
- Zoomdata - 大数据分析平台。

数据可视化

- Airpal - PrestoDB 的 Web UI。
- AnyChart - 一套灵活的 JavaScript (HTML5) 库，可满足您的所有数据可视化需求。
- Arbor - 一个使用 web workers 和 jQuery 创建的图可视化库。
- Banana - 可视化存储在 Solr 中的日志和带时间戳的数据，是 Kibana 的一部分。
- Bloomery - Impala 的 Web UI。
- Bokeh - 一个 Python 交互式可视化库，支持现代化 Web 浏览器，提供非常完美的展示功能。
- C3 - 基于 D3 的可重用图表库
- CartoDB - 开源的云上地理空间数据库，允许存储和可视化 web 上的数据。使用 CartoDB 可以快速创建基于地图的可视化效果。
- chartd - 响应式、视网膜兼容图表，仅需要一个 img 标签。
- Chart.js - 一套开源、简单、干净并且有吸引力的基于 HTML5 技术的 JavaScript 图表工具。
- Chartist.js - 非常简单而且实用的 JavaScript 前端图表生成器。
- Crossfilter - 一个 JavaScript 库，用于在 JavaScript 中制作交互式的仪表板，可以与 dc.js 、d3.js 一起工作。
- Cubism - 用于时间序列可视化的 JavaScript 库。
- Cytoscape - 一个专注于网络可视化和分析的开源软件。
- DC.js - 一个用于网页作图、生成互动图形的 JavaScript 函数库。

- D3 - 目前最流行的数据可视化库之一，小型，灵活，高效的数据可视化库，用来创建和操作基于数据的交互式文档。
- D3.compose - 由可重复使用的图表和组件组成复杂的、数据驱动的可视化文件。
- D3Plus - d3.js 的一组相当强大的可重用图表和样式。
- DevExtreme React Chart - 基于高性能插件的 React 图表，用于 Bootstrap 和 Material Design。
- Echarts - 一款由百度前端技术部开发的，基于Javascript 的数据可视化图表库，提供直观，生动，可交互，可个性化定制的数据可视化图表。
- Envisionjs - 一个基于 HTML5 技术的数据可视化库
- FnordMetric - 一个开源的 Web 应用，可用于创建实时仪表板，方便可视化任何数据。
- Frappe Charts - 一个受 Github 启发的轻量级 SVG 图表库，它不依赖任何类库和框架。
- Freeboard - 让用户创建他们自己的用来监控物联网部署的仪表盘，该代码在 GitHub上免费提供，你可以通过这些仪表板展示跟踪空气质量、住宅电器、酿酒情况和实时环境条件变化。
- Gephi - 一款开源免费跨平台基于 JVM 的网络分析领域的数据可视化处理软件
- Google Charts - 一种交互式 Web 服务，可根据用户提供的数据创建图形图表
- Grafana - 一个跨平台的开源的度量分析和可视化工具，可以通过将采集的数据查询然后可视化的展示，并及时通知。
- Graphite - 一款开源的监控绘图工具。
- Highcharts - 兼容 IE6+、完美支持移动端、图表类型丰富、方便快捷的 HTML5 交互性图表库。
- IPython - 一种基于 Python 的交互式解释器。相较于原生的 Python Shell，IPython 提供了更为强大的编辑和交互功能。
- Kibana - Elasticsearch 的开源数据可视化插件。
- Lumify - 开源大数据分析可视化平台。
- Matplotlib - Python 编程语言及其数值数学扩展包 NumPy 的可视化操作界面。
- Metricsgraphic.js - 一个建立在 D3 基础上，为可视化和时间序列化的数据而优化的库。
- NVD3 - d3.js 的图表组件。
- Peity - 渐进式 SVG 条形图，折线图和饼图。
- Plot.ly - Plotly 为个人和协作提供在线图形，分析和统计工具，以及 Python，R，MATLAB，Perl，Julia，Arduino 和 REST 的科学图形库。
- Plotly.js 一个开源的交互式 JavaScript 图形库，建立在 d3.js 和 webgl 之上，并支持 20 多种类型的交互式图表。
- Recline - 简单而强大的库，可以使用纯 Javascript 和 HTML 构建数据应用程序。
- Redash - 查询和可视化数据的开源平台。
- ReCharts - 一个基于React组件的可组合图表库。
- Shiny - R 的 Web 应用程序框架。
- Sigma.js - 专门用于图形绘制的 JavaScript 库。
- Superset - 由 Airbnb 开发并开源一个数据探索和可视化平台，设计用来提供直观的，可视化的，交互式的分析体验。
- Vega - 一个可视化的语法。
- Zeppelin - 一个基于 Web 的 notebook，提供交互数据分析和可视化。
- Zing Charts - 一个功能强大的 JavaScript 图表。

物联网和传感器数据

- Apache Edgent (Incubating) -
一种编程模型和具有微内核风格的运行时，可嵌入到网关和小型的物联网设备中。
- Azure IoT Hub - 托管服务，支持 IoT 设备与 Azure 之间的双向通信。
- TempoIQ - 基于云计算的传感器分析。.
- 2lemetry - 物联网平台。
- Pubnub - 数据流网络。
- ThingWorx - 可用于查找数据来源，使数据与情境相关，合成数据，同时协调流程，以提供强大的Web、移动和AR 体验的平台。
- IFTTT -
一个新生的网络服务平台，通过其他不同平台的条件来决定是否执行下一条命令。
- Evrything- 使产品智能化。
- NetLytics - 用于在Spark上处理网络数据的分析平台。

有趣的阅读材料

- Big Data Benchmark - Redshift，Hive，Shark，Impala 和 Stiger/Tez的基准。
- NoSQL Comparison - Cassandra，MongoDB，CouchDB，Redis，Riak，HBase，Couchbase，Neo4j，Hypertable，ElasticSearch，Accumulo，VoltDB 和 Scalaris 的比较。
- Monitoring Kafka performance - 监视 Apache Kafka
的指南，包括度量收集的本地方法。
- Monitoring Hadoop performance - 监视 Hadoop 的指南，概述了 Hadoop
体系结构以及度量收集的本机方法。
- Monitoring Cassandra performance - 监控 Cassandra
的指南，包括度量收集的本地方法。

有趣的论文

2015 - 2016

- 2015 - Facebook - One Trillion Edges: Graph Processing at Facebook-Scale.

2013 - 2014

- 2014 - Stanford - Mining of Massive Datasets.
- 2013 - AMPLab - Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices.
- 2013 - AMPLab - MLbase: A Distributed Machine-learning System.
- 2013 - AMPLab - Shark: SQL and Rich Analytics at Scale.
- 2013 - AMPLab - GraphX: A Resilient Distributed Graph System on Spark.
- 2013 - Google - HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality Estimation Algorithm.
- 2013 - Microsoft - Scalable Progressive Analytics on Big Data in the Cloud.
- 2013 - Metamarkets - Druid: A Real-time Analytical Data Store.
- 2013 - Google - Online, Asynchronous Schema Change in F1.
- 2013 - Google - F1: A Distributed SQL Database That Scales.

- 2013 - Google - MillWheel: Fault-Tolerant Stream Processing at Internet Scale.
- 2013 - Facebook - Scuba: Diving into Data at Facebook.
- 2013 - Facebook - Unicorn: A System for Searching the Social Graph.
- 2013 - Facebook - Scaling Memcache at Facebook.

2011 - 2012

- 2012 - Twitter - The Unified Logging Infrastructure for Data Analytics at Twitter.
- 2012 - AMPLab - Blink and It's Done: Interactive Queries on Very Large Data.
- 2012 - AMPLab - Fast and Interactive Analytics over Hadoop Data with Spark.
- 2012 - AMPLab - Shark: Fast Data Analysis Using Coarse-grained Distributed Memory.
- 2012 - Microsoft - Paxos Replicated State Machines as the Basis of a High-Performance Data Store.
- 2012 - Microsoft - Paxos Made Parallel.
- 2012 - AMPLab - BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data.
- 2012 - Google - Processing a trillion cells per mouse click.
- 2012 - Google - Spanner: Google's Globally-Distributed Database.
- 2011 - AMPLab - Scarlett: Coping with Skewed Popularity Content in MapReduce Clusters.
- 2011 - AMPLab - Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center.
- 2011 - Google - Megastore: Providing Scalable, Highly Available Storage for Interactive Services.

2001 - 2010

- 2010 - Facebook - Finding a needle in Haystack: Facebook's photo storage.
- 2010 - AMPLab - Spark: Cluster Computing with Working Sets.
- 2010 - Google - Pregel: A System for Large-Scale Graph Processing.
- 2010 - Google - Large-scale Incremental Processing Using Distributed Transactions and Noti □ cations base of Percolator and Caffeine.
- 2010 - Google - Dremel: Interactive Analysis of Web-Scale Datasets.
- 2010 - Yahoo - S4: Distributed Stream Computing Platform.
- 2009 - HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads.
- 2008 - AMPLab - Chukwa: A large-scale monitoring system.
- 2007 - Amazon - Dynamo: Amazon's Highly Available Key-value Store.
- 2006 - Google - The Chubby lock service for loosely-coupled distributed systems.
- 2006 - Google - Bigtable: A Distributed Storage System for Structured Data.
- 2004 - Google - MapReduce: Simplied Data Processing on Large Clusters.
- 2003 - Google - The Google File System.

视频

- Spark in Motion - Spark in Motion 教你如何使用 Spark 进行批处理和流数据分析。

图书

Streaming

- Data Science at Scale with Python and Dask - Data Science at Scale with Python and Dask teaches you how to build distributed data projects that can handle huge amounts of data.
- Streaming Data - Streaming Data introduces the concepts and requirements of streaming and real-time data systems.
- Storm Applied - Storm Applied is a practical guide to using Apache Storm for the real-world tasks associated with processing and analyzing real-time data streams.
- Fundamentals of Stream Processing: Application Design, Systems, and Analytics - This comprehensive, hands-on guide combining the fundamental building blocks and emerging research in stream processing is ideal for application designers, system builders, analytic developers, as well as students and researchers in the field.
- Stream Data Processing: A Quality of Service Perspective - Presents a new paradigm suitable for stream and complex event processing.
- Unified Log Processing - Unified Log Processing is a practical guide to implementing a unified log of event streams (Kafka or Kinesis) in your business
- Kafka Streams in Action - Kafka Streams in Action teaches you everything you need to know to implement stream processing on data flowing into your Kafka platform, allowing you to focus on getting more from your data without sacrificing time or effort.
- Big Data - Big Data teaches you to build big data systems using an architecture that takes advantage of clustered hardware along with new tools designed specifically to capture and analyze web-scale data.
- Spark in Action & Spark in Action 2nd Ed. - Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. Fully updated for Spark 2.0.
- Kafka in Action - Kafka in Action is a fast-paced introduction to every aspect of working with Kafka you need to really reap its benefits.
- Fusion in Action - Fusion in Action teaches you to build a full-featured data analytics pipeline, including document and data search and distributed data clustering.
- Reactive Data Handling - Reactive Data Handling is a collection of five hand-picked chapters, selected by Manuel Bernhardt, that introduce you to building reactive applications capable of handling real-time processing with large data loads--free eBook!

Distributed systems

- Distributed Systems for fun and profit – 分布式系统理论。包括时间、顺序、副本等。

Graph Based approach

- Graph-Powered Machine Learning - Alessandro Negro，结合图论和模型改进机器学习项目

Data Visualization

- The beauty of data visualization
- Designing Data Visualizations with Noah Iliinsky
- Hans Rosling's 200 Countries, 200 Years, 4 Minutes
- Ice Bucket Challenge Data Visualization

本文翻译自：Awesome Big Data