
Awesome Data Engineering awesome

A curated list of awesome things related to Data Engineering.

Contents

- Databases
- Data Comparison
- Data Ingestion
- File System
- Serialization format
- Stream Processing
- Batch Processing
- Charts and Dashboards
- Workflow
- Data Lake Management
- ELK Elastic Logstash Kibana
- Docker
- Datasets
 - Realtime
 - Data Dumps
- Monitoring
 - Prometheus
- Profiling
 - Data Profiler
- Testing
- Community
 - Forums
 - Conferences
 - Podcasts

Databases

- Relational

-
- RQLite - Replicated SQLite using the Raft consensus protocol.
 - MySQL - The world's most popular open source database.
 - * TiDB - TiDB is a distributed NewSQL database compatible with MySQL protocol.
 - * Percona XtraBackup - Percona XtraBackup is a free, open source, complete online backup solution for all versions of Percona Server, MySQL® and MariaDB®.
 - * mysql_utils - Pinterest MySQL Management Tools.
 - MariaDB - An enhanced, drop-in replacement for MySQL.
 - PostgreSQL - The world's most advanced open source database.
 - Amazon RDS - Amazon RDS makes it easy to set up, operate, and scale a relational database in the cloud.
 - Crate.IO - Scalable SQL database with the NOSQL goodies.
- Key-Value
 - Redis - An open source, BSD licensed, advanced key-value cache and store.
 - Riak - A distributed database designed to deliver maximum data availability by distributing data across multiple servers.
 - AWS DynamoDB - A fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale.
 - HyperDex - HyperDex is a scalable, searchable key-value store. Deprecated.
 - SSDB - A high performance NoSQL database supporting many data structures, an alternative to Redis.
 - Kyoto Tycoon - Kyoto Tycoon is a lightweight network server on top of the Kyoto Cabinet key-value database, built for high-performance and concurrency.
 - IonDB - A key-value store for microcontroller and IoT applications.
 - Column
 - Cassandra - The right choice when you need scalability and high availability without compromising performance.
 - * Cassandra Calculator - This simple form allows you to try out different values for your Apache Cassandra cluster and see what the impact is for your application.
 - * CCM - A script to easily create and destroy an Apache Cassandra cluster on localhost.
 - * ScyllaDB - NoSQL data store using the seastar framework, compatible with Apache Cassandra.
 - HBase - The Hadoop database, a distributed, scalable, big data store.
 - AWS Redshift - A fast, fully managed, petabyte-scale data warehouse that makes it simple and cost-effective to analyze all your data using your existing business intelligence tools.
 - FiloDB - Distributed. Columnar. Versioned. Streaming. SQL.
 - Vertica - Distributed, MPP columnar database with extensive analytics SQL.

-
- ClickHouse - Distributed columnar DBMS for OLAP. SQL.
 - Document
 - MongoDB - An open-source, document database designed for ease of development and scaling.
 - ★ Percona Server for MongoDB - Percona Server for MongoDB® is a free, enhanced, fully compatible, open source, drop-in replacement for the MongoDB® Community Edition that includes enterprise-grade features and functionality.
 - ★ MemDB - Distributed Transactional In-Memory Database (based on MongoDB).
 - Elasticsearch - Search & Analyze Data in Real Time.
 - Couchbase - The highest performing NoSQL distributed database.
 - RethinkDB - The open-source database for the realtime web.
 - RavenDB - Fully Transactional NoSQL Document Database.
 - Graph
 - Neo4j - The world's leading graph database.
 - OrientDB - 2nd Generation Distributed Graph Database with the flexibility of Documents in one product with an Open Source commercial friendly license.
 - ArangoDB - A distributed free and open-source database with a flexible data model for documents, graphs, and key-values.
 - Titan - A scalable graph database optimized for storing and querying graphs containing hundreds of billions of vertices and edges distributed across a multi-machine cluster.
 - FlockDB - A distributed, fault-tolerant graph database by Twitter. Deprecated.
 - Distributed
 - DAtomic - The fully transactional, cloud-ready, distributed database.
 - Apache Geode - An open source, distributed, in-memory database for scale-out applications.
 - Gaffer - A large-scale graph database.
 - Timeseries
 - InfluxDB - Scalable datastore for metrics, events, and real-time analytics.
 - OpenTSDB - A scalable, distributed Time Series Database.
 - QuestDB - A relational column-oriented database designed for real-time analytics on time series and event data.
 - kairoddb - Fast scalable time series database.
 - Heroic - A scalable time series database based on Cassandra and Elasticsearch, by Spotify.
 - Druid - Column oriented distributed data store ideal for powering interactive applications.

-
- Riak-TS - Riak TS is the only enterprise-grade NoSQL time series database optimized specifically for IoT and Time Series data.
 - Akumuli - Akumuli is a numeric time-series database. It can be used to capture, store and process time-series data in real-time. The word “akumuli” can be translated from esperanto as “accumulate”.
 - Rhombus - A time-series object store for Cassandra that handles all the complexity of building wide row indexes.
 - Dalmatiner DB - Fast distributed metrics database.
 - Blueflood - A distributed system designed to ingest and process time series data.
 - Timely - Timely is a time series database application that provides secure access to time series data based on Accumulo and Grafana.
- Other
 - Tarantool - Tarantool is an in-memory database and application server.
 - GreenPlum - The Greenplum Database (GPDB) - An advanced, fully featured, open source data warehouse. It provides powerful and rapid analytics on petabyte scale data volumes.
 - cayley - An open-source graph database. Google.
 - Snappydata - SnappyData: OLTP + OLAP Database built on Apache Spark.
 - TimescaleDB - Built as an extension on top of PostgreSQL, TimescaleDB is a time-series SQL database providing fast analytics, scalability, with automated data management on a proven storage engine.

Data Comparison

- datacompy - DataComPy is a Python library that facilitates the comparison of two DataFrames in pandas, Polars, Spark and more. The library goes beyond basic equality checks by providing detailed insights into discrepancies at both row and column levels.

Data Ingestion

- Kafka - Publish-subscribe messaging rethought as a distributed commit log.
 - BottledWater - Change data capture from PostgreSQL into Kafka. Deprecated.
 - kafkat - Simplified command-line administration for Kafka brokers.
 - kafkacat - Generic command line non-JVM Apache Kafka producer and consumer.
 - pg-kafka - A PostgreSQL extension to produce messages to Apache Kafka.
 - librdkafka - The Apache Kafka C/C++ library.
 - kafka-docker - Kafka in Docker.

-
- kafka-manager - A tool for managing Apache Kafka.
 - kafka-node - Node.js client for Apache Kafka 0.8.
 - Secor - Pinterest's Kafka to S3 distributed consumer.
 - Kafka-logger - Kafka-winston logger for Node.js from uber.
 - AWS Kinesis - A fully managed, cloud-based service for real-time data processing over large, distributed data streams.
 - RabbitMQ - Robust messaging for applications.
 - dlt - A fast&simple pipeline building library for python data devs, runs in notebooks, cloud functions, airflow, etc.
 - FluentD - An open source data collector for unified logging layer.
 - Embulk - An open source bulk data loader that helps data transfer between various databases, storages, file formats, and cloud services.
 - Apache Sqoop - A tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.
 - Heka - Data Acquisition and Processing Made Easy. Deprecated.
 - Gobblin - Universal data ingestion framework for Hadoop from LinkedIn.
 - Nakadi - Nakadi is an open source event messaging platform that provides a REST API on top of Kafka-like queues.
 - Pravega - Pravega provides a new storage abstraction - a stream - for continuous and unbounded data.
 - Apache Pulsar - Apache Pulsar is an open-source distributed pub-sub messaging system.
 - AWS Data Wrangler - Utility belt to handle data on AWS.
 - Airbyte - Open-source data integration for modern data teams.
 - Sling - Sling is CLI data integration tool specialized in moving data between databases, as well as storage systems.

File System

- HDFS - A distributed file system designed to run on commodity hardware.
 - Snakebite - A pure python HDFS client.
- AWS S3 - Object storage built to retrieve any amount of data from anywhere.
 - smart_open - Utils for streaming large files (S3, HDFS, gzip, bz2).
- Alluxio - Alluxio is a memory-centric distributed storage system enabling reliable data sharing at memory-speed across cluster frameworks, such as Spark and MapReduce.
- CEPH - Ceph is a unified, distributed storage system designed for excellent performance, reliability and scalability.

-
- OrangeFS - Orange File System is a branch of the Parallel Virtual File System.
 - SnackFS - SnackFS is our bite-sized, lightweight HDFS compatible FileSystem built over Cassandra.
 - GlusterFS - Gluster Filesystem.
 - XtremFS - Fault-tolerant distributed file system for all storage needs.
 - SeaweedFS - Seaweed-FS is a simple and highly scalable distributed file system. There are two objectives: to store billions of files! to serve the files fast! Instead of supporting full POSIX file system semantics, Seaweed-FS choose to implement only a key~file mapping. Similar to the word “NoSQL”, you can call it as “NoFS”.
 - S3QL - S3QL is a file system that stores all its data online using storage services like Google Storage, Amazon S3, or OpenStack.
 - LizardFS - LizardFS Software Defined Storage is a distributed, parallel, scalable, fault-tolerant, Geo-Redundant and highly available file system.

Serialization format

- Apache Avro - Apache Avro™ is a data serialization system.
- Apache Parquet - Apache Parquet is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model or programming language.
 - Snappy - A fast compressor/decompressor. Used with Parquet.
 - PigZ - A parallel implementation of gzip for modern multi-processor, multi-core machines.
- Apache ORC - The smallest, fastest columnar storage for Hadoop workloads.
- Apache Thrift - The Apache Thrift software framework, for scalable cross-language services development.
- ProtoBuf - Protocol Buffers - Google’s data interchange format.
- SequenceFile - SequenceFile is a flat file consisting of binary key/value pairs. It is extensively used in MapReduce as input/output formats.
- Kryo - Kryo is a fast and efficient object graph serialization framework for Java.

Stream Processing

- Apache Beam - Apache Beam is a unified programming model that implements both batch and streaming data processing jobs that run on many execution engines.
- Spark Streaming - Spark Streaming makes it easy to build scalable fault-tolerant streaming applications.

-
- Apache Flink - Apache Flink is a streaming dataflow engine that provides data distribution, communication, and fault tolerance for distributed computations over data streams.
 - Apache Storm - Apache Storm is a free and open source distributed realtime computation system.
 - Apache Samza - Apache Samza is a distributed stream processing framework.
 - Apache NiFi - An easy to use, powerful, and reliable system to process and distribute data.
 - Apache Hudi - An open source framework for managing storage for real time processing, one of the most interesting feature is the Upsert.
 - VoltDB - VoltDb is an ACID-compliant RDBMS which uses a shared nothing architecture.
 - PipelineDB - The Streaming SQL Database.
 - Spring Cloud Dataflow - Streaming and tasks execution between Spring Boot apps.
 - Bonobo - Bonobo is a data-processing toolkit for python 3.5+.
 - Robinhood's Faust - Forever scalable event processing & in-memory durable K/V store as a library with asyncio & static typing.
 - HStreamDB - The streaming database built for IoT data storage and real-time processing.
 - Kuiper - An edge lightweight IoT data analytics/streaming software implemented by Golang, and it can be run at all kinds of resource-constrained edge devices.
 - Zilla -- An API gateway built for event-driven architectures and streaming that supports standard protocols such as HTTP, SSE, gRPC, MQTT and the native Kafka protocol.

Batch Processing

- Hadoop MapReduce - Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) - in-parallel on large clusters (thousands of nodes) - of commodity hardware in a reliable, fault-tolerant manner.
- Spark - A multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.
 - Spark Packages - A community index of packages for Apache Spark.
 - Deep Spark - Connecting Apache Spark with different data stores. Deprecated.
 - Spark RDD API Examples - Examples by Zhen He.
 - Livy - The REST Spark Server.
 - Delight - A free & cross platform monitoring tool (Spark UI / Spark History Server alternative).
- AWS EMR - A web service that makes it easy to quickly and cost-effectively process vast amounts of data.

-
- Data Mechanics - A cloud-based platform deployed on Kubernetes making Apache Spark more developer-friendly and cost-effective.
 - Tez - An application framework which allows for a complex directed-acyclic-graph of tasks for processing data.
 - Bistro - A light-weight engine for general-purpose data processing including both batch and stream analytics. It is based on a novel unique data model, which represents data via *functions* and processes data via *columns operations* as opposed to having only set operations in conventional approaches like MapReduce or SQL.
 - Batch ML
 - H2O - Fast scalable machine learning API for smarter applications.
 - Mahout - An environment for quickly creating scalable performant machine learning applications.
 - Spark MLlib - Spark's scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives.
 - Batch Graph
 - GraphLab Create - A machine learning platform that enables data scientists and app developers to easily create intelligent apps at scale.
 - Giraph - An iterative graph processing system built for high scalability.
 - Spark GraphX - Apache Spark's API for graphs and graph-parallel computation.
 - Batch SQL
 - Presto - A distributed SQL query engine designed to query large data sets distributed over one or more heterogeneous data sources.
 - Hive - Data warehouse software facilitates querying and managing large datasets residing in distributed storage.
 - ★ Hivemall - Scalable machine learning library for Hive/Hadoop.
 - ★ PyHive - Python interface to Hive and Presto.
 - Drill - Schema-free SQL Query Engine for Hadoop, NoSQL and Cloud Storage.

Charts and Dashboards

- Highcharts - A charting library written in pure JavaScript, offering an easy way of adding interactive charts to your web site or web application.
- ZingChart - Fast JavaScript charts for any data set.

-
- C3.js - D3-based reusable chart library.
 - D3.js - A JavaScript library for manipulating documents based on data.
 - D3Plus - D3's simpler, easier to use cousin. Mostly predefined templates that you can just plug data in.
 - SmoothieCharts - A JavaScript Charting Library for Streaming Data.
 - PyXley - Python helpers for building dashboards using Flask and React.
 - Plotly - Flask, JS, and CSS boilerplate for interactive, web-based visualization apps in Python.
 - Apache Superset - Apache Superset (incubating) - A modern, enterprise-ready business intelligence web application.
 - Redash - Make Your Company Data Driven. Connect to any data source, easily visualize and share your data.
 - Metabase - Metabase is the easy, open source way for everyone in your company to ask questions and learn from data.
 - PyQtGraph - PyQtGraph is a pure-python graphics and GUI library built on PyQt4 / PySide and numpy. It is intended for use in mathematics / scientific / engineering applications.

Workflow

- Luigi - Luigi is a Python module that helps you build complex pipelines of batch jobs.
 - CronQ - An application cron-like system. Used w/Luige. Deprecated.
- Cascading - Java based application development platform.
- Airflow - Airflow is a system to programmatically author, schedule and monitor data pipelines.
- Azkaban - Azkaban is a batch workflow job scheduler created at LinkedIn to run Hadoop jobs. Azkaban resolves the ordering through job dependencies and provides an easy to use web user interface to maintain and track your workflows.
- Oozie - Oozie is a workflow scheduler system to manage Apache Hadoop jobs.
- Pinball - DAG based workflow manager. Job flows are defined programmatically in Python. Support output passing between jobs.
- Dagster - Dagster is an open-source Python library for building data applications.
- Kedro - Kedro is a framework that makes it easy to build robust and scalable data pipelines by providing uniform project templates, data abstraction, configuration and pipeline assembly.
- Dataform - An open-source framework and web based IDE to manage datasets and their dependencies. SQLX extends your existing SQL warehouse dialect to add features that support dependency management, testing, documentation and more.
- Census - A reverse-ETL tool that let you sync data from your cloud data warehouse to SaaS applications like Salesforce, Marketo, HubSpot, Zendesk, etc. No engineering favors required—just

SQL.

- dbt - A command line tool that enables data analysts and engineers to transform data in their warehouses more effectively.
- RudderStack - A warehouse-first Customer Data Platform that enables you to collect data from every application, website and SaaS platform, and then activate it in your warehouse and business tools.
- PACE - An open source framework that allows you to enforce agreements on how data should be accessed, used, and transformed, regardless of the data platform (Snowflake, BigQuery, DataBricks, etc.)
- Prefect - Prefect is an orchestration and observability platform. With it, developers can rapidly build and scale resilient code, and triage disruptions effortlessly.
- Multiwoven - The open-source reverse ETL, data activation platform for modern data teams.
- SuprSend - Create automated workflows and logic using API's for your notification service. Add templates, batching, preferences, inapp inbox with workflows to trigger notifications directly from your data warehouse.

Data Lake Management

- lakeFS - lakeFS is an open source platform that delivers resilience and manageability to object-storage based data lakes.
- Project Nessie - Project Nessie is a Transactional Catalog for Data Lakes with Git-like semantics. Works with Apache Iceberg tables.

ELK Elastic Logstash Kibana

- docker-logstash - A highly configurable logstash (1.4.4) - docker image running Elasticsearch (1.7.0) - and Kibana (3.1.2).
- elasticsearch-jdbc - JDBC importer for Elasticsearch.
- ZomboDB - Postgres Extension that allows creating an index backed by Elasticsearch.

Docker

- Gockerize - Package golang service into minimal docker containers.
- Flocker - Easily manage Docker containers & their data.
- Rancher - RancherOS is a 20mb Linux distro that runs the entire OS as Docker containers.
- Kontena - Application Containers for Masses.
- Weave - Weaving Docker containers into applications.

-
- Zodiac - A lightweight tool for easy deployment and rollback of dockerized applications.
 - cAdvisor - Analyzes resource usage and performance characteristics of running containers.
 - Micro S3 persistence - Docker microservice for saving/restoring volume data to S3.
 - Rocker-compose - Docker composition tool with idempotency features for deploying apps composed of multiple containers. Deprecated.
 - Nomad - Nomad is a cluster manager, designed for both long lived services and short lived batch processing workloads.
 - ImageLayers - Vizualize docker images and the layers that compose them.

Datasets

Realtime

- Twitter Realtime - The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data.
- Eventsim - Event data simulator. Generates a stream of pseudo-random events from a set of users, designed to simulate web traffic.
- Reddit - Real-time data is available including comments, submissions and links posted to reddit.

Data Dumps

- GitHub Archive - GitHub's public timeline since 2011, updated every hour.
- Common Crawl - Open source repository of web crawl data.
- Wikipedia - Wikipedia's complete copy of all wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available.

Monitoring

Prometheus

- Prometheus.io - An open-source service monitoring system and time series database.
- HAProxy Exporter - Simple server that scrapes HAProxy stats and exports them via HTTP for Prometheus consumption.

Profiling

Data Profiler

- Data Profiler - The DataProfiler is a Python library designed to make data analysis, monitoring, and sensitive data detection easy.

Testing

- Grai - A data catalog tool that integrates into your CI system exposing downstream impact testing of data changes. These tests prevent data changes which might break data pipelines or BI dashboards from making it to production.
- DQOps - An open-source data quality platform for the whole data platform lifecycle from profiling new data sources to applying full automation of data quality monitoring.

Community

Forums

- /r/dataengineering - News, tips and background on Data Engineering.
- /r/etl - Subreddit focused on ETL.

Conferences

- Data Council - Data Council is the first technical conference that bridges the gap between data scientists, data engineers and data analysts.

Podcasts

- Data Engineering Podcast - The show about modern data infrastructure.
- The Data Stack Show - A show where they talk to data engineers, analysts, and data scientists about their experience around building and maintaining data infrastructure, delivering data and data products, and driving better outcomes across their businesses with data.