Data Stack Overview version 1.0



Scrapers/Crawlers Data

<u>Scrapy</u>	Opensource	Powerfull Python Scraping Framework
Nutch	Opensource	High Scalability Crawler
<u>PhantomJS</u>	Lib	Headless Browser perfect for scraping
X-ray	Lib	Very fast and simple scraper
<u>ImportIO</u>	Saas	Webscraping as a Service
<u>Morph</u>	Saas	Multi Lang Scraper Platform

Website Data

<u>Piwik</u>	Opensource/ Saas	On Premise alternative with Saas option
Google Analytics	Saas	General purpose analytics
Google 360	Saas	Enterprise level
<u>Mixpanel</u>	Saas	Crossdevice
<u>Kissmetrics</u>	Saas	Ecommerce oriented
<u>Woopra</u>	Saas	Ecommerce oriented
<u>Adobe</u> <u>Analytics</u>	Saas	Enterprise level
Fox Metrics	Saas	Ecommerce oriented
<u>GoSquared</u>	Saas	Focus on CRM

IoT Data/Analytics

<u>Particle</u>	\$	IoT with Wifi
Raspberry Pi	\$	Standard
<u>UbiDots</u>	Saas	Data Collection and Analysis
<u>Thingspeak</u>	Saas	Data Collection and Analysis
<u>Xively</u>	Saas	Data Collection and Analysis

Social Media Data

<u>Gnip</u>	Saas	Raw Social Media Data
<u>Datasift</u>	Saas	Raw Social Media Data
<u>Spinn3r</u>	Saas	Social Media Crawler



Logging/Analysis

<u>Graylog</u>	Opensource	Opensource Logging Service
Logstash	Opensource	From ELK Stack
<u>Fluentd</u>	Opensource	Open Source Data collector
Graphite/ graphana	Opensource	Good greping support, ELK stack competitor
Loggly	Saas	Like Graylog but paid
<u>Splunk</u>	Saas	Industry strength log analysis
<u>Sumologic</u>	Saas	Cloud hosted log analysis

Message Queue/streams

<u>Kafka</u>	Opensource	PubSub Stream
RabbitMQ	Opensource	Message Queue
<u>Flink</u>	Opensource	Stream Analytics
<u>Storm</u>	Opensource	Distributed realtime computation system.
<u>Flume</u>	Opensource	Log Data Streams
<u>Airflow</u>	Opensource	Data Pipelines
<u>Celery</u>	Opensource	Distributed Task Queue
<u>Impala</u>	Opensource	Realtime Stream Query
<u>Druid</u>	Opensource/ Saas	Streams and Column DB
Google PubSub	Saas	PubSub Stream
<u>Kinesis</u>	Saas	PubSub Stream
<u>Dataflow</u>	Saas	Orchestration
Amazon SQS	Saas	Simple Queue Service in a Cloud

Data Cleansing/Transforming

<u>Openrefine</u>	Opensource	Easy Data Cleansing
Wrangler	Opensource	Academic Data Cleansing
<u>Nifi</u>	Opensource	Process and Distribute Data
<u>SQOOP</u>	Opensource	Hadoop <-> SQL
<u>Trifacta</u>	\$	Commercial version of wrangler
TalenD	\$	Datalake Transformation



Database

<u>SQLITE</u>	Oponsourco	Filohasad/Prototypa DR
	Opensource	Filebased/Prototype DB
<u>Postgres</u>	Opensource	SQL
<u>MYSQL</u>	Saas/GPL	SQL
<u>OracleDB</u>	Saas	SQL
<u>MariaDB</u>	Opensource	SQL
<u>Neo4j</u>	Opensource	Graph / NoSQL
<u>Tinkerpop</u>	Opensource	Graph / NoSQL
<u>RocksDb</u>	Opensource	Key/Value
<u>Redis</u>	Opensource	Key/Value
<u>LeveIDB</u>	Opensource	Key/Value
<u>MongoDB</u>	Opensource	NoSQL
<u>Riak</u>	Saas/GPL	NoSQL
<u>CouchDB</u>	Opensource	NoSQL
<u>Couchbase</u>	Opensource	NoSQL
<u>Google</u> <u>Bigtable</u>	Saas	NoSQL
<u>Aerospike</u>	Opensource	NoSQL
<u>Presto</u>	Saas/GPL	MPP
<u>Teradata</u>	SaasSaas	MPP
<u>Snowflake</u>	Saas	DWH
<u>LucidDb</u>	Opensource	DWH
<u>MonetDB</u>	Opensource	Column Oriented
<u>Hypertable</u>	Opensource	Column Oriented
SAP Hana	Saas	Column Oriented
<u>HP Vertica</u>	Saas	Column Oriented
<u>MLDB</u>	Opensource	DB with ML
<u>Pytables</u>	Opensource	Hierarchical Data
<u>BlinkDB</u>	Opensource	Sampling DB
KX DB	Opensource	Fast Column oriented with own query language
<u>Axibase</u>	Opensource	Time Series Database with Viz
HDF5 Data Format	Opensource	High density format for data

	Obyotom	
<u>Hadoop</u>	Opensource	Distributed Storage and Computation / Map Reduce
<u>Hbase</u>	Opensource	Distributed Database/Filesystem
<u>Cassandra</u>	Opensource	Distributed database
<u>Hive</u>	Opensource	DWH for Hadoop
<u>Oozie</u>	Opensource	Workflow Scheduler
<u>Openstack</u>	Opensource	Open source software for creating private and public clouds.
<u>Mahout</u>	Opensource	Distributed realtime computation system.
<u>Spark</u>	Opensource	General engine for large-scale data processing.
Redshift	Saas	Amazon DWH Solution
<u>S3</u>	Saas	Distributed Storage
<u>Cloudera</u>	Opensource/ Saas	Haddop Ecosystem
<u>Hortonworks</u>	Opensource/ Saas	Hadoop on premise or cloud

<u>Elastic</u>	Opensource	General flexible fast solution
<u>Lucene</u>	Opensource	Mature Search engine
<u>Solr</u>	Opensource	Builds on Lucene
<u>Lunar</u>	Opensource	Inbrowser search engine
<u>Compass</u>	Opensource	Java search engine
<u>Sphinx</u>	Opensource	Java search engine
<u>Xapian</u>	Opensource	C search engine
<u>VoltDB</u>	Opensource	Fast inmem column db
<u>Lemur/Indri</u>	Opensource	Java Search Engine
<u>Websolr</u>	Saas	Hosted Solr
Searchify	Saas	Hosted Search
<u>Algolia</u>	Saas	Hosted Search with focus on integration
<u>Searchblox</u>	Saas	Hosted Search
<u>Bonsai</u>	Saas	Elastic Saas
Searchly	Saas	Elastic Saas
Azure Search	Saas	Microsoft Saas search
<u>Cloudsearch</u>	Saas	Amazon Saas search
<u>Exasol</u>	Saas	Industry strength fast inmem db

Analysis ML

Deep learning frameworks		
<u>Tensorflow</u>	Opensource	Googles Lowlevel Deeplearning framework
Deeplearning4 J	Opensource	Highlevel Deeplearning Java
<u>Torch</u>	Opensource	General Deeplearn lua
<u>Theano</u>	Opensource	Lowlevel Deeplearning on GPUs
<u>MXNet</u>	Opensource	Lightweight deep learning framework
<u>TFLearn</u>	Lib	Highlevel Deeplearning (now part of tensorflow)
<u>Keras</u>	Lib	Highlevel Deeplearning
<u>Caffee</u>	Lib	Highlevel Deeplearning
<u>Brainstorm</u>	Lib	Sucessor of pybrain
<u>Pybrain</u>	Lib	Python Machine Learning Library with focus on NN
Google ML	Saas	Hosted Deeplearning Infrastructure

State software and nackages

Stats software and packages		
<u>lpython</u>	Opensource	Ipython Notebooks
Rattle	Opensource	Graphical UI for R models
Rapidminer	Opensource	Java Data Modeling and Prediction
<u>Weka</u>	Opensource	ML industry strength toolkit in java
<u>Orange</u>	Opensource	Data Modeling and prediction
<u>R</u>	Opensource	Big opensource stats framefork mainly used in academia
<u>Julia</u>	Opensource	Lang for Scientific Comp.
<u>Octave</u>	Opensource	Like matlab but open source
<u>Pandas</u>	Lib	Easy-to-use data structures and data analysis tools in python
<u>SPSS</u>	Saas	IBMs industrial strength stats in java
<u>SAS</u>	Saas	Proprietary stats modeling
<u>Matlab</u>	Saas	Industrial strength numerical computing environment

General ML (focus on python)

Scikit Learn	Lib	A Python module for machine learning built on top of SciPy.
<u>REP</u>	Lib	Ipython Scikit Extension
<u>XBoost</u>	Lib	Python bindings for eXtreme Gradient Boosting (Tree) Librar
Scipy	Lib	Python framework for mathematics, science, and engineering.
Numpy	Lib	A fundamental package for scientific computing with Python.
<u>Statsmodels</u>	Lib	Statistical modeling and econometrics in Python.
<u>AutoML</u>	Lib	Automated machine learning pipelines for analytics and
<u>Aerosolve</u>	Opensource	A human friendly machine learning library by Airbnb.
<u>NetworkX</u>	Lib	A high-productivity software for complex networks
<u>Pandas</u>	Lib	Time series handling in pandas.

Computer Vision

	SciKit Image	Lib	A collection of algorithms for image processing in Python
	Google CV	Saas	Saas for Image recognition
	SimpleCV	Lib	Framework that gives access to OpenCV.
	<u>Clarifai</u>	Saas	Saas for Image recognition
	Microsoft CV	Saas	Saas for Image recognition
	Diffbot	Saas	Saas for Image recognition

Natural Language Processing

	Python NLTK	Lib	Lib for text processing in Python
	<u>Spacy</u>	Saas	Industrial strength NLP with Python
	<u>Pattern</u>	Lib	Pattern recognition package in Go lang
	<u>Lexalitics</u>	Saas/Lib	NLP as a Service
	<u>OpenCalais</u>	Saas	Named Entities from Text
	Cortical	Saas	Retina: an Saas performing complex NLP operations
	<u>CoreNLP</u>	Lib	Integrated NLP Toolkit
	<u>LDAjs</u>	Lib	Topic Modeling
	<u>FuzzyWuzzy</u>	Lib	Fuzzy Stringmatching in Ruby
	Texacy	Lib	Highlevel Spacy
	<u>Natural.js</u>	Lib	General Language utilities for node
	<u>Gensim</u>	Lib	Topic Modeling
	<u>TM</u>	Lib	Text mining for R
	Google NLP	Saas	NLP Deep learning Models from google

Visualisation Dashboards Apps

For R

General Visualisation

<u>Matplotlib</u>	Lib	General Purpose Lib
<u>D3</u>	Lib	General very exensible JS lib
<u>Chartjs</u>	Lib	Easy to use js lib
<u>Bookeh</u>	Lib	Sophisticated python lib
<u>HighCharts</u>	Lib	Industry strength js viz
NVD3	Lib	Highlevel D3
<u>Seaborn</u>	Lib	Sophisticated python lib
<u>Sigmajs</u>	Lib	Network Viz in js
<u>Datamaps</u>	Lib	SVG Map visualization
<u>Odyssey</u>	Lib	Stories visualization in js
<u>Cubism</u>	Lib	Time series viz in js
<u>Gatherplot</u>	Lib	Gatherplots in js
<u>Glue</u>	Lib	Linked Data viz in python
<u>Altair</u>	Lib	Vegalite for Python
<u>Vega</u>	Lib	Declarative Viz in js
<u>Gephi</u>	Opensource	Network Visualization
<u>DC</u>	Opensource	Multi Dimensional Charting
<u>Crossfilter</u>	Opensource	Multi Dimensional Filtering
<u>Zeppelin</u>	Opensource	Crosslanguage Interactive Viz Notebook
<u>Folium</u>	Opensource	Python Map visualization with leafletjs
<u>Leafletjs</u>	Opensource	js lib for map visualization
Google Charts	Saas	Golden classic
Raw	Saas	Rather esoteric stuff but cool
<u>Timeline</u>	Saas/ Opensource	Timeline viz in ja
Shiny	Saas/ Opensource	Visualization of R models
<u>Variancecharts</u>	Saas/ Opensource	Like GGplot for js
<u>Datawrapper</u>	Saas/ Opensource	Focus on embedding and responsiveness in js
<u>Amcharts</u>	Saas/ Opensource	Fancy js viz
Anychart	Saas/ Opensource	Robust js lib

Dashboards

ReDash	Opensource/ Saas	Dashboard with Analytics
<u>Vida</u>	Opensource/ Saas	Flexible dashboard
Screenly	Opensource/ Saas	Manage multiple dashboards
Plotly	Opensource/ Saas	Dashboards wit a lot of Bindings
Dashing	Lib	Sinatra based framework with realtime
<u>Freeboard</u>	Saas	Dashboard for IoT
<u>Keen</u>	Saas	Dashboard and Analytics
Geckoboard	Saas	Dashboards for TVs
Telemetry	Saas	Dashboards for TVs
Klipfolio	Saas	Easy to use dashboard

Bots Frameworks

	<u>Botpress</u>	Opensource	Open Source Framwork for Bots
	<u>Aiva</u>	Lib	General-purpose virtual assistar for developers
	<u>P-Brain</u>	Lib	Natural Language for Bots
	<u>Howdy</u>	Lib	Open Source Bot Kit
	<u>WitAI</u>	Saas	Natural Language for Bots
	<u>Api.ai</u>	Saas	Natural Language for Bots

Reports / Emails



Business Intelligence

		J
<u>Kibana</u>	Opensource	Fast Moving BI on top of elastic
<u>Airpal</u>	Opensource	BI from airbnb
<u>OpenMining</u>	Opensource	BI in Python Notebooks
<u>Blazer</u>	Opensource	Simple SQL to Graph
<u>Metabase</u>	Saas/ Opensource	Easy to fork and extend
<u>Qlik</u>	Saas/ Opensource	Highly customizatble BI
IBM Cognos	Saas	Integrated BI from IBM
Oracle BI	Saas	Integrated BI from Oracle
SAP BI	Saas	Integrated BI from SAP
<u>PowerBI</u>	Saas	Integrated BI from Microsoft
<u>Tableau</u>	Saas	Industry standard visualization/ analytics
<u>ChartIO</u>	Saas	Easy to use BI
<u>Sisense</u>	Saas	Easy to use inmem BI
<u>Pentaho</u>	Saas	Integrated BI
<u>Looker</u>	Saas	Fancy BI platform

Data Viz and Analytics

Business Intelligence (continued)

<u>Cartodb</u>	Saas	BI for Maps Data
<u>Quoble</u>	Saas	Integrated BlgData Analytics
Micro stratetegy	Saas	Integrated BI
<u>Opentext</u>	Saas	BI focused on Oracle stack
Logi Analytics	Saas	BI with aspect on Integration
Yellow Fin	Saas	Easy to start BI
<u>Domo</u>	Saas	BI with focus on realtime
Good Data	Saas	Fortune 500 BI
<u>Dundas</u>	Saas	BI with focus on Flexibility
Information- builders	Saas	BI with desktop UI
<u>Jedox</u>	Saas	Excel like BI
Pyramid Analytics	Saas	Classic BI
<u>Dimensional</u> <u>Insight</u>	Saas	Integrated BI
<u>Tibco</u>	Saas	Integrated BI
<u>Birst</u>	Saas	Integrated BI

Business Intelligence (continuted)

<u>1010Data</u>	Saas	Integrated BI
Adaptive Insights	Saas	BI with focus on Finance
Rosslyn Analytics	Saas	Automated BI
<u>PeriscopeData</u>	Saas	BI with focus on SQL queries
Excel Powerpivot	Saas	Pivot Tables in Excel
SAS Visual Analytics	Saas	Like Tableau from SAS

<u>Arcadia Data</u>	Saas	BI on Hadoop
<u>Datameer</u>	Saas	BI on Hadoop
Kyvos Insights	Saas	BI on Hadoop
<u>Attivio</u>	Saas	BI on Hadoop
<u>Zoomdata</u>	Saas	BI on Hadoop

Data Science Platforms

	1	
<u>Knime</u>	Saas/ Opensource	BI for Datascientists
<u>Datarobot</u>	Saas	Integrated Datascience
<u>Dominolab</u>	Saas	Integrated Datascience
<u>Yhat</u>	Saas	Model deployment
<u>Modeanalytics</u>	Saas	Integrated Datascience
<u>Dataiku</u>	Saas	Integrated Datascience
<u>Continuum</u>	Saas	Anaconda Datascience Suite
<u>Snowplow</u>	Saas	Integrated Event analysis
<u>BigMI</u>	Saas	Machine Learning Platform

Integrated Datascience



BI with quite a lot of connectors