The Starter Guide

For the Modern Data Stack

Why is there a “Modern” Stack?

The way we use data in our lives has drastically changed over the last decade.

And it shouldn’t be a surprise that the way companies manage it has needed to evolve as well.

We’ve gone from only needing one or two main tools in an an-hoc or scheduled “ETL” process...

To a more decoupled “ELT” approach with *many* specialized tools and real-time automation.

The good news is...

We now have many more options. *(hurray!)*

This means better solutions for each component and more advanced functionality.

The bad news is...

We now have many more options. *(now I’m stressed out)*

Each day there seems to be a new tool to learn and it’s hard to mentally put it all together.

This guide will help you cut through the noise and feel more confident with how it all fits together in the modern stack.

The Starter Guide

For the Modern Data Stack

What's Inside?

The first visual will give you a high level overview of a typical modern ELT process.

It purposely *does not include* any tools so you can instead focus on the high level components.

Otherwise you'll immediately get lost in all of the options and lose sight of the big picture.

Use this visual to help you identify *high-level topics* to learn more about.

The second visual adds logos of common products/tools that you will often see used.

Use this visual to identify *individual tools or platforms* to learn more about.

An important thing to note is that there are MANY options for the modern stack.

Including many other great tools outside of what I've listed.

It would become too overwhelming (and not productive) to list every possible option.

The last page is a breakdown of each of the logos used along with a description.

Use this page as a *reference guide* and/or to do further research.

The Starter Guide

For the Modern Data Stack

A final thought before you dive in...

This architecture can feel overwhelming, especially if you're just getting started.

But just know that most companies will typically only use one or two from each category.

And others may skip certain parts completely.

You don't need to master every single tool or language to have a successful career.

Use this guide to improve your awareness and guide your journey through this wild world of the modern data stack.

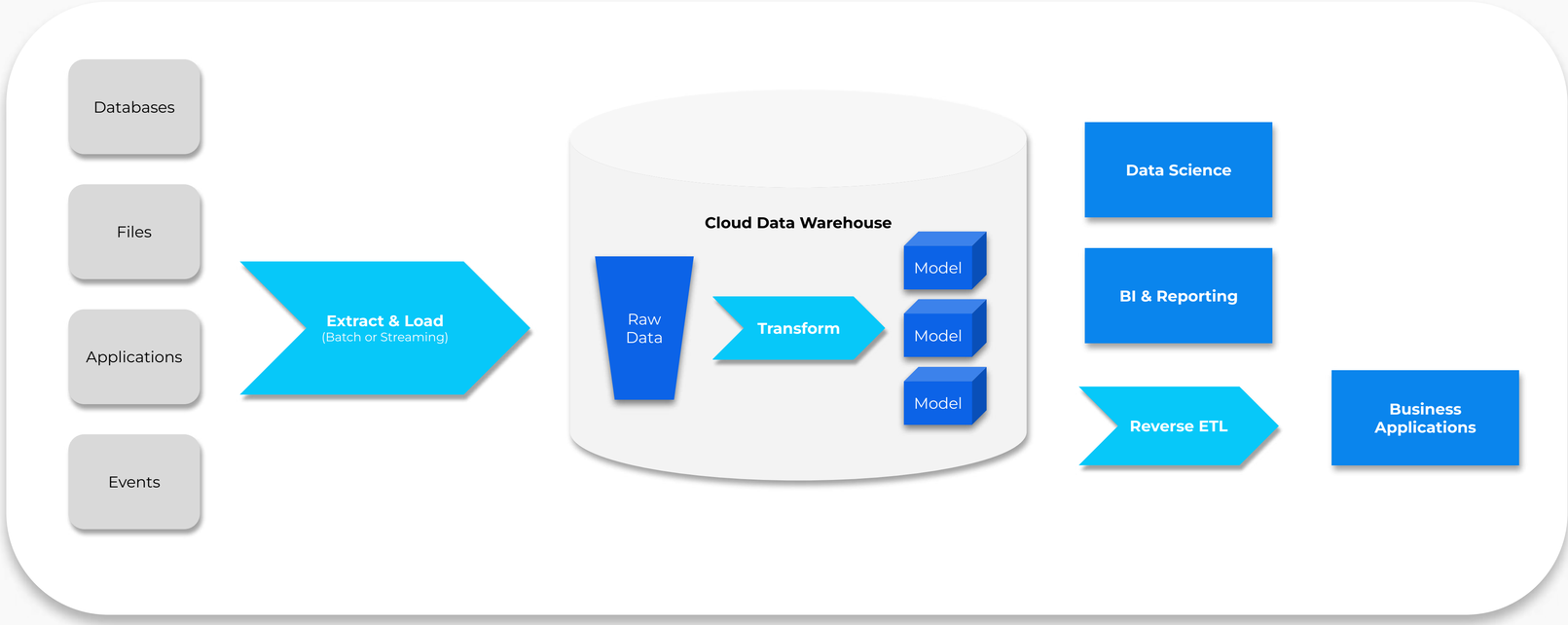
Cheers,
Mike

The Starter Guide

For the Modern Data Stack

Languages & Data Types

SQL | Python | Jinja | JSON/XML | YAML | DAX | Bash / PowerShell



Cloud Providers | Data Quality | Version Control & CI/CD | Infrastructure as Code (IaC) | Containerization | Container Orchestration | Scheduling & Task Orchestration

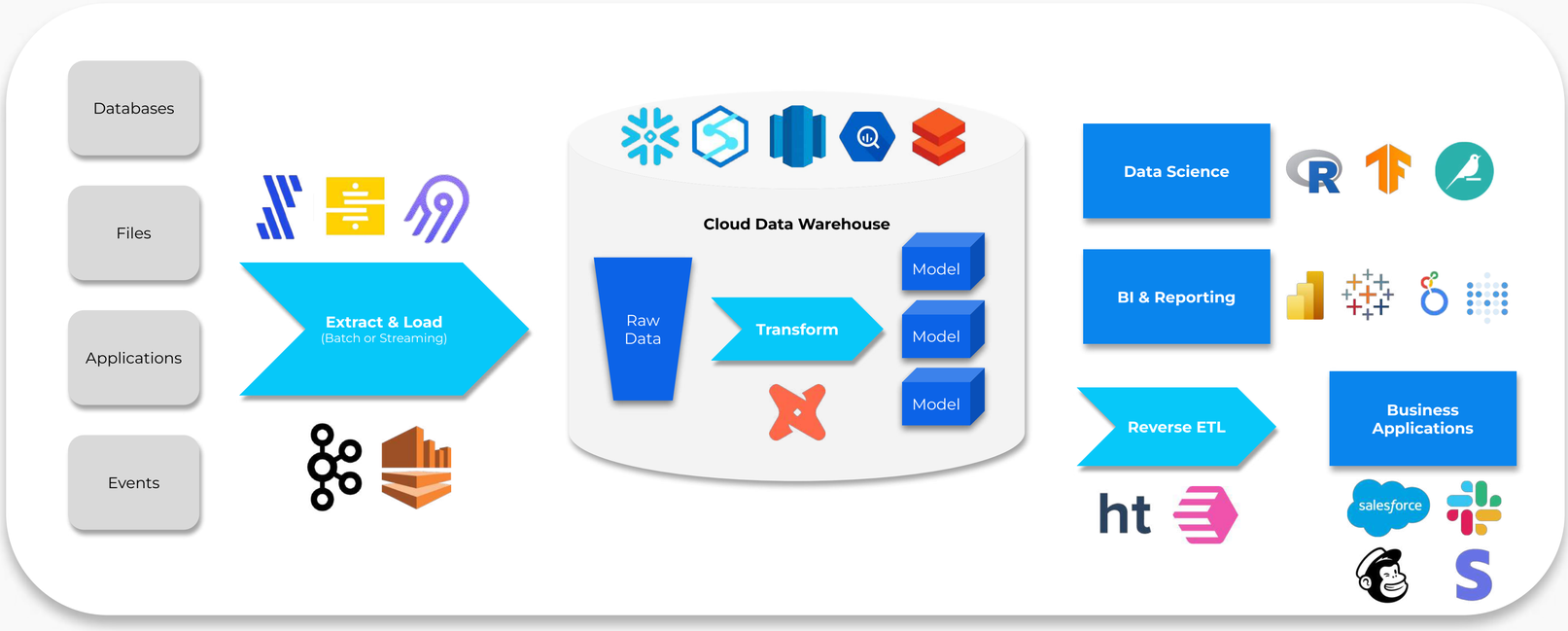
Surrounding Components

The Starter Guide

For the Modern Data Stack

Languages & Data Types

SQL | Python | Jinja | JSON/XML | YAML | DAX | Bash / PowerShell



Surrounding Components

Cloud Providers | Data Quality | Version Control & CI/CD | Infrastructure as Code (IaC) | Containerization | Container Orchestration | Scheduling & Task Orchestration

The Starter Guide

For the Modern Data Stack

Extract & Load

Fivetran
https://www.fivetran.com/
Fivetran is a data integration service for companies to extract, transform and load data from different sources into data warehouses.

Stitch
https://www.stitchdata.com/
Stitch is a cloud-first, open source platform for rapidly moving data. A simple, powerful ETL service, Stitch connects to all your data sources and replicates that data to a destination of your choosing.

Airbyte
https://airbyte.com/
Airbyte is an open-source data integration platform to build ELT pipelines. Consolidate your data in your data warehouses, lakes and databases.

Apache Kafka
https://kafka.apache.org/
Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications.

Amazon Kinesis
https://aws.amazon.com/kinesis/
Collect streaming data, create a real-time data pipeline, and perform real-time clickstream analytics, log analytics, event analytics, and IoT analytics.

Cloud Data Warehouse

Snowflake
https://www.snowflake.com/
The Snowflake data warehouse is a cloud-based tool that supplies companies with flexible and scalable storage while simultaneously hosting solutions for BI.

Azure Synapse Analytics
https://azure.microsoft.com/en-us/services/synapse-analytics/
Azure Synapse is an enterprise analytics service that brings together data warehousing and other data technologies such as Spark and data pipelines.

Amazon Redshift
https://aws.amazon.com/redshift/
Amazon Redshift uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes, using AWS-designed hardware and machine learning to deliver the best price performance at any scale.

Google BigQuery
https://cloud.google.com/bigquery
BigQuery is a fully-managed, serverless data warehouse that enables scalable analysis over petabytes of data. It also has built-in machine learning capabilities.

Databricks
https://databricks.com/
Built on Apache Spark, it provides a data processing engine that many companies use with a Data Warehouse. It can also be used as a Data Lakehouse by using Databricks Delta Lake and Delta Engine.

Transform

data-build-tool (dbt)
https://www.getdbt.com/
dbt™ is a transformation workflow that lets teams quickly and collaboratively deploy analytics code following software engineering best practices like modularity, portability, CI/CD, and documentation.

Cloud Providers

Amazon Web Services (AWS)
https://aws.amazon.com/
AWS is the world's most comprehensive and broadly adopted cloud platform, offering over 200 fully featured services from data centers globally.

Microsoft Azure
https://azure.microsoft.com/
The Azure cloud platform is more than 200 products and cloud services. Build, run, and manage applications across multiple clouds, on-premises, and at the edge, with the tools and frameworks of your choice.

Google Cloud Platform (GCP)
https://cloud.google.com/
GCP is a set of Computing, Networking, Storage, Big Data, Machine Learning and Management services provided by Google that runs on the same Cloud infrastructure that Google uses internally for its end-user products.

IBM Cloud
https://www.ibm.com/cloud
IBM cloud computing is a set of cloud computing services for business offered by the information technology company IBM

Oracle Cloud Infrastructure (OCI)
https://www.oracle.com/cloud/
Oracle Cloud is a cloud computing service offered by Oracle Corporation providing servers, storage, network, applications and services through a global network of Oracle Corporation managed data centers.

Data Quality

Great Expectations
https://greatexpectations.io/
Great Expectations is a shared, open standard for data quality. It helps data teams eliminate pipeline debt, through data testing, documentation, and profiling.

SQL Fluff
https://www.sqlfluff.com/
SQL Fluff is an extensible and modular linter designed to help you write good SQL and catch errors and bad SQL before it hits your database.

Apache Griffin
https://griffin.apache.org/
Apache Griffin is an open source Data Quality solution for Big Data, which supports both batch and streaming mode. It offers an unified process to measure your data quality from different perspectives.

Version Control & CI/CD

GitHub
https://github.com/
Millions of developers and companies build, ship, and maintain their software on GitHub—the largest and most advanced development platform in the world.

GitLab
https://about.gitlab.com/
GitLab is The DevOps platform that empowers organizations to maximize the overall return on software development by delivering software faster and efficiently, while strengthening security and compliance.

Bitbucket
https://bitbucket.org/
With best-in-class Jira integration, and built-in CI/CD, Bitbucket Cloud is the native Git tool in Atlassian's Open DevOps solution.

BI & Reporting

Power BI
https://powerbi.microsoft.com/
Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence.

Tableau
https://www.tableau.com/
Tableau is a visual analytics platform transforming the way we use data to solve problems—empowering people and organizations to make the most of their data.

Looker
https://www.looker.com/
Looker is a business intelligence software and big data analytics platform that helps you explore, analyze and share real-time business analytics easily.

Metabase
https://www.metabase.com/
Metabase is an open source business intelligence tool. It lets you ask questions about your data, and displays answers in formats that make sense, whether that's a bar chart or a detailed table.

Reverse ETL

Census
https://www.getcensus.com/
Census is the leading reverse ETL tool delivering operational analytics to hundreds of the world's most data-driven companies. Sync data from your warehouse into all your business tools.

Hightouch
https://hightouch.io/
The leading Data Activation platform. Hightouch syncs customer data from your warehouse to the tools that your business teams rely on.

Data Science

R
https://www.r-project.org/
R is an open source programming language and free software that is used by data scientists, data miners and statisticians for developing statistical software and data analysis.

TensorFlow
https://www.tensorflow.org/
TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.

Dataiku
https://www.dataiku.com/
Dataiku is one central solution for the design, deployment, and management of AI applications. It's designed for teams who want to deliver advanced analytics using the latest techniques at big data scale.

Business Applications

Salesforce
https://www.salesforce.com/
Salesforce is a company that makes cloud-based software designed to help businesses find more prospects, close more deals, and wow customers with amazing service.

Slack
https://slack.com/
Slack is the collaboration hub that brings the right people, information, and tools together to get work done. Millions of people around the world use Slack to connect their teams, unify their systems, and drive their business forward.

Mailchimp
https://mailchimp.com/
Mailchimp is a marketing automation platform and email marketing service for managing mailing lists and creating email marketing campaigns to send to customers

Stripe
https://stripe.com/
Stripe is a suite of APIs powering online payment processing and commerce solutions for internet businesses of all sizes. Accept payments and scale faster.

Infrastructure as Code (IaC)

Terraform
https://www.terraform.io/
HashiCorp Terraform is an infrastructure as code tool that lets you define both cloud and on-prem resources in human-readable configuration files that you can version, reuse, and share.

Red Hat Ansible
https://www.ansible.com/
Ansible is an open source community project sponsored by Red Hat, it's the simplest way to automate IT. It is the only automation language that can be used across entire IT teams from administrators to developers and managers.

Containerization

Docker
https://www.docker.com/
Docker helps conquer the complexity of app development. It simplifies and accelerates development workflows with an integrated dev pipeline and through the consolidation of application components.

Container Orchestration

Kubernetes
https://kubernetes.io/
Kubernetes is a portable, extensible, open source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation.

Red Hat Openshift
https://www.redhat.com/en/technologies/cloud-computing/openshift
OpenShift is an enterprise-ready Kubernetes container platform built for an open hybrid cloud strategy. It provides a consistent application platform to manage hybrid cloud, multicloud, and edge deployments.

Scheduling & Task Orchestration

Apache Airflow
https://airflow.apache.org/
Airflow is an open-source platform created by the community to programmatically author, schedule and monitor workflows.

Prefect
https://www.prefect.io/
Prefect is a new workflow management system, designed for modern infrastructure and powered by the open-source Prefect Core workflow engine.

Jenkins
https://www.jenkins.io/
Jenkins is an open source automation server which enables developers around the world to reliably build, test, and deploy their software.

Dagster
https://dagster.io/
Dagster is an orchestration platform for the development, production, and observation of data assets. Develop and test locally, then deploy anywhere.

Luigi
https://luigi.readthedocs.io/en/stable/#
Luigi is a Python package that helps you build complex pipelines of batch jobs. It handles dependency resolution, workflow management, visualization, handling failures, command line integration, and much more.