Data Science and Engineering Today

(and the Impact on the World)

panoply.io

Table of Contents

Brilliant computer and data scientists have spent the past few decades sharing their wisdom with the masses and turning that wisdom into products that have resulted in the market that we enjoy today. Back when the first hard drive was sold or when the first digital database management system went into production, who would have thought that the world would need a myriad of database options?

But the world of the mid-19th century and the world of today are vastly different places. In the span of just a few decades, mankind has harnessed the atom, invested the transistor, and unleashed unto the world a vast communications network that has reshaped business, society, and the way we live.

And it's all powered by data. Whereas access to data and powerful information technology systems used to be considered a competitive business differentiator, today, access to such services is a prerequisite for business success. Today, businesses are looking for new and innovative ways to leverage data to get a jump on the competition, and all sorts of new tools are being developed to help them down this path.

Of course, the traditional relational database continues to power mission critical systems where data availability and consistency are key. Increasingly, though, availability and consistency are sometimes optional, and tradeoffs are being made in order to be able to scale systems to the levels that are necessary to support today's massive data sets and data variability.

State of NoSQL

NoSQL began life at companies like Google and Facebook, where traditional database systems were not able to keep pace with service growth. In fact, the popular Cassandra tool was actually developed at Facebook to power the company's search capabilities. Although Facebook has since moved to another product for search, Cassandra is still used in some other services. Moreover, a few years ago, Facebook turned Cassandra over to the open source community where it is thriving.

It makes sense that these large webscale companies would require specialized tools to support the sheer scale of their services. However, one of the interesting things that tends to happen with these kinds of tools is that they eventually make their way to the enterprise and beyond. Rather than being relegated to supporting the search function for a social network, the tools can be used to help a "regular" business undertake complex data analysis or develop large distributed services of their own.

The need to manage massive amounts of data can't be overstated. In this paper, you'll be exposed to some of today's popular products and services and will learn how data is playing a bigger role in your life than you might think.

Common Implementations

So, NoSQL isn't really new, per se, but its popularity is exploding. In fact, there are hundreds of potential NoSQL database tools available for your consumption, according to **http://nosql-database.org/**. This site provides a complete list of all of the tools at your disposal and also provides a description of each one, and, in many cases, how it works.

Not all of these tools can possible be described in this paper, so we'll focus on three tools that have become very popular. In this section, you'll get a brief introduction to the tool and will be presented with some key facts about each one.

Apache HBase

Maintained by the Apache Foundation, HBase is a column-oriented key-value store and is a part of the wider Hadoop ecosystem, serving as its database. HBase is an open source implementation of Bigtable, Google's massively scalable NoSQL database. HBase has become a popular tool among cloud-scale companies, such as Facebook, which uses HBase to power its massive messaging platform.

If you look at HBase in the context of Eric Brewer's CAP Theorem, HBase falls in the CP category, meaning that it is consistent and has partition tolerance.

From the project documentation, HBase carries the following features:

- **Strongly consistent reads/writes**

HBase is not an "eventually consistent" DataStore. This makes it very suitable for tasks such as high-speed counter aggregation.

- **Automatic sharding**

HBase tables are distributed on the cluster via regions, and regions are automatically split and re-distributed as your data grows.

HBase is a distributed database, although the project prefers to consider itself a"data store" than "database". The reason: HBase lacks many of the features found in an RDBMS, such as typed columns, secondary indexes, triggers, and advanced query languages.

MongoDB

MongoDB is a free, open source document store NoSQL database. As is the case with many open source projects, MongoDB is also available in a commercial version, called MongoDB Enterprise Advanced. MongoDB is available via two distribution methods.

On-premises. Assuming that you have a server environment and operating systems that MongoDB supports, you can deploy the solution in your private data center. MongoDB supports most modern x86 platforms and operating systems, some ARM-based platforms that run Ubuntu 16.04, some PPCLE systems, and even IBM zSystems running Ubuntu's s390x ported operating system.

MongoDB Atlas is the cloud version. Available as a database-as-a-service offering in the cloud, MongoDB Atlas enables you to get a full-featured document database without having to deploy anything locally. Just sign up for an account and get started.

On the data front, MongoDB stores data using what is called BSON, which stands for Binary JavaScript Object Notation. According to **http://bsonspec.org/**:

> "BSON is a binary-encoded serialization of JSON-like documents. Like JSON, BSON supports the embedding of documents and arrays within other documents and arrays. BSON also contains extensions that allow representation of data types that are not part of the JSON spec. For example, BSON has a Date type and a BinData type, providing it with more flexibility than JSON alone."

Source: http://bsonspec.org

Cassandra

Like HBase, Cassandra is another Apache project. A column-oriented database, Cassandra was first created at Facebook to support the inbox search function. For projects that require large scale deployments that span multiple data centers and that require top notch speed, Cassandra is a great choice.

Although it's a NoSQL database, Cassandra includes Cassandra Query Language (CQL), which very much resembles SQL. This can make Cassandra a bit easier to adopt for those that are used to more traditional SQL approaches.

Transformational Infrastructure Technologies

The various NoSQL tools on the market are just a part of the equation. In order to work, tools need a place to run. There have been two major developments over the past decade that are worthy of discussion as they have done more to advance data science and engineering than just about anything else.

Cloud

You've probably heard of the cloud. Amazon, Azure and the like. While the term "the cloud" has been the unfortunate victim of overzealous tech marketing folks, the principles that make the cloud what it is are sound.

Let's explore.

Prior to the cloud, most organizations ran their own infrastructure environment, either on-premises in their own data centers or in rented colocation facilities. Regardless of where it lived, all of the equipment was generally owned or leased by the company. When the time came for a new application or business service to be deployed, IT sprang into action and bought a bunch of hardware to support that new initiative.

Let's go through a scenario in which Acme, Inc. intends to deploy a new data reporting and analysis solution. The process begins with the CEO, W.E. Coyote, telling his CIO, Rex Runner, that the business wants a new data analytics platform and specifies the product that will be used. Under a traditional IT model, Mr. Runner would request that his infrastructure team begin a comprehensive requirements analysis and then procure, install, and configure all the hardware and software that the intended solution requires. The infrastructure components would involve servers, storage, and networking equipment and, on the software side, there would be operating systems and possibly hypervisors to deploy, not to mention the analytics software itself. Because Mr. Coyote considers the data analytics platform mission

critical, Rex Runner would also ensure that his IT team deploy a highly available infrastructure, which might include running component both locally and at a backup data center.

This is the way that IT used to be managed and, in many cases, still is. However, there are a number of challenges inherent in this model:

- **Forecasting.**

The IT crystal ball isn't perfect. IT may end up buying either too much or too little hardware. After all, it's entirely possible that the business will end up using the new analytics tool in ways that IT didn't anticipate. This could have an impact on the infrastructure.

- **Cost**

The business has to buy all of the hardware and software up front as a part of a huge capital purchase. It can take years to actually recoup these expenses, if it happens at all.

- **Flexibility**

If IT does "miss" on forecasting, it can sometimes be a challenge—technically and financially—to upgrade infrastructure mid-cycle. Most equipment in IT is placed on some kind of a replacement cycle, which is often based on depreciation. If there are unexpected bumps along the way, there can be a lot of disruption.

There's nothing inherently wrong with this model. It's been used for years, but, with the advent of cloud and thanks to the aforementioned rise of the Internet, there is a way to get services without the downsides mentioned above.

Organizations around the world are turning to cloud operational models in order to streamline their IT operations and to help control costs. There are a number of potential upsides to cloud adoption, including:

- **Cost.**

With the cloud, an organization essentially rents IT resources on a pay-as-you-go basis from a cloud provider. This obviates the need to pay huge upfront capital costs when new applications are required.

• Time.

In the Acme, Inc. example outlined above, it may have taken the IT team weeks or even months to fully respond to the new business need. During that time, the company would have been spending money, but not seeing an immediate return and they would be missing out on the benefits expected from the new application.

• Flexibility.

The beauty of the cloud is that there are practically unlimited resources available for consumption. You can buy as much or as little as you want or need. So, in the event that you need more resources to run that new analytics application, you simply aquire additional resources from the cloud provider. You will see an associated increase in your monthly invoice from the provider, but you won't have to spend a lot of new money to buy new infrastructure due to inaccurate initial planning or changes in how an application is operated.

There are several different models that you can consider when it comes to IT and cloud operations. Each is outlined in Figure X-X. In the On-Prem column, you will notice that everything is managed and supported by the local IT team. This is the traditional IT model. To the right, you will see three primary cloud operating models:

• Infrastructure-as-a-Service (IaaS).

Under this model, the provider makes available the underlying infrastructure, including networks, storage, compute resources, and virtualization technology. However, your staff still has responsibility for configuring these resources and you still continue to manage security, databases, and applications.

• Platform-as-a-Service (PaaS).

PaaS environments provide infrastructure and an application development platform. They often include the ability to automate and deploy applications. PaaS provides operating systems, databases, middleware, tools and services, leaving the customer to manage just the application and data layers.

• Software-as-a-Service (SaaS).

This is the simplest level of cloud-based service that you can consume. Under this model, the provider controls everything so your local IT team can focus on other

needs. The provider makes available an application layer interface only for specific configuration items. In general, you do not need to worry about any underlying services except those that may extend the service to help it to integrate with your on-premises infrastructure.

Operating Models	On-Prem	IaaS	PaaS	SaaS
Application	Local IT	Local IT	Local IT	Provider
Database	Local IT	Local IT	Local IT	Provider
Middleware	Local IT	Local IT	Provider	Provider
Operating System	Local IT	Local IT	Provider	Provider
Hypervisor	Local IT	Provider	Provider	Provider
Physical Server	Local IT	Provider	Provider	Provider
Storage	Local IT	Provider	Provider	Provider
Network	Local IT	Provider	Provider	Provider

Figure 3-1. Comparing IT Deployment Models.

For data engineers, the cloud is a cornucopia of opportunity. Consider the example discussed early. Suppose Acme Inc.'s CIO, Rex Runner, was told that he needed to deploy a cloud-based data analytics tool, such as the one provided by Panoply. Rather than weeks or months, getting up and running can take minutes or hours. There would be no infrastructure to set up, so business decision makers could begin gaining value from the solution immediately.

Today's IT environments are a mix of local infrastructure and these public cloud offerings, a scenario which is often termed as hybrid cloud. CIOs and other decision makers have the opportunity to consider all aspects of a new workload deployment and then make a decision as to where to run it and how to enable the end goal. Sometimes that will mean buying a service like Panoply or Microsoft Office 365; sometimes that will mean adopting Amazon Web Services (AWS) and deploying

workloads there. And sometimes that will mean deploying infrastructure and services locally. Again, there is no wrong answer and different companies will have different needs and, as a result, will choose different solutions.

FLASH STORAGE

Let's revisit Acme, Inc. Except, rather than cloud, Acme has made the decision that they're going to deploy their new analytics application locally. Rex Runner, Acme's

CIO, has to choose a storage environment on which to operate the new application. Let's assume that Rex chooses to use all spinning disk for this need rather than all-flash or hybrid (a combination of flash and spinning disk). Here are the challenges that he will face:

- **Performance.**

As you know, when it comes to sheer speed, spinning disk is far inferior to flash. That spinning disk environment will provide only a small fraction of the performance that a flash—or even a hybrid storage—solution will allow.

- **Disk sprawl.**

With disk only, to get more performance you have to add more spindles (disks). So, if you need things to be faster, you'll have to add far more disks than you may actually need from a capacity perspective.

- **Power draw.**

With normal usage patterns, flash storage consumes far less power than spinning disk. And, consider that, to get more performance, you need more disks. That further increases the power draw. Oh, and more disks equals more heat, which means you need more power for cooling. With spinning disk, your electrical costs will likely be far higher than with flash.

Should Rex choose to go the all-flash or hybrid storage route, he will find that he has a number of benefits:

- **Lower operational costs.**

For the reasons described above, Rex will likely have lower ongoing maintenance costs.

- **Far more performance.**

Rex will have IOPS to spare! Flash storage provides users with massive performance capabilities.

• Maybe, increased capacity.

Most flash storage solution on the market today have the ability to reduce duplicate data and compress data so that it consumes less overall capacity. Depending on the kind of data you're working with, this data reduction can be significant.

Why is flash so much faster than disk? Well, that's because it works at the speed of light.

NAND FLASH TECHNICAL BASICS

Common NAND flash-based storage systems work by trapping electrical charges in a floating gate/transistor combination (cell). The data value is then derived by determining voltage values in these individual cells. There are different kinds of flash media available on the market. Single level cell (SLC) was the original type. It's the most expensive kind around, but it's also the fastest. SLC stores just a single bit of data per cell. Multi-level cell (MLC) came after SLC, and doubles its capacity by allowing the media to store two bits of data per cell. This capacity comes at the expense of performance and durability. MLC is somewhat slower than SLC and it wears out faster than SLC. And then there's triple level cell (TLC) media. TLC stores, as you may have guessed, three bits per cell, and is somewhat slower than MLC and also less durable. As time goes on, however, it's becoming more robust and more common since it's still much faster than spinning disk.

In flash, "durability" refers to the number of times that each cell can be erased and reprogrammed before it wears out. Early on, there were major concerns around whether or not flash media could stay alive long enough to make sense. Fortunately, manufacturers have implemented technologies that effectively extend the life of the flash media and those initial concerns around durability have been largely unnecessary.

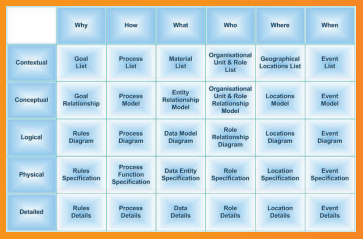
In recent years, to continue to increase the capacity of flash storage without linearly increasing costs, manufacturers have started creating what is known as 3D NAND. As with the common wisdom in construction, it's cheaper to build up rather than out, and that's similar mentality for 3D NAND.

Perhaps the most significant benefit of flash storage is its raw speed; it doesn't even compare to spinning disk. Whereas a single spinning disk may be able to support a few hundred (at most) I/O operations per second (IOPS), a single flash disk can

support tens of thousands of IOPS, although the exact value is directly dependent on the characteristics of the data stream.

THE FLASH BOTTOM LINE

When it comes to data engineering, flash is a game changer. The technology enables fast reading and writing for even the largest data sets. And, even though cloud and flash were presented as separate options in this paper, in reality, you can get both. Cloud environments are just huge data centers managed by someone else. They can have flash storage as well, and many cloud providers make this high-performance storage available as an option.



John Zackman (1936-)

Enterprise Architecture via the Zackman Framework

John Zackman was a founding development of IBM's Business Systems Planning service, an effort dedicated to helping businesses make sense of how to support the intersection of data, business processes, business strategies, and organizational hierarchy. This was an early effort at business process engineering with some focus on how technology can be leveraged to improve the business.

Zackman went on to create his own overall enterprise architecture framework, which he named the Zackman Framework. The goal of the framework is to describe the enterprise. By combining basic communication (why, how, what, who, where, when) with audience perspective, the model seeks to define how the enterprise operates. With this understanding can come improvement. While the model can be general in nature, it's generally focused on the technology function.

From the Drawing Board to Real Life

Perhaps you've spent a good chunk of your career managing, for example, networks, and you think that the data revolution hasn't impacted you. Well… it has. In fact, most people in the developed world are impacted by data every day and that's not likely to change. In fact, as computers get more powerful, and as people become even more interconnected, the amount of data being captured and leveraged will grow exponentially. In this section, we'll briefly cover some of the data-rich technologies and services on the market today.

ARTIFICIAL INTELLIGENCE (AI)

The human brain is a data-driven organism. Everything humans do is based on our experiences and knowledge, both of which are forms of data. The actions we take are built on this data.

In order for artificial intelligence to work, vast quantities of data must be brought to bear. The AI needs to have sufficient background and training in order to be able to continuously learn and adapt to new situations and new questions.

Many look to tools such as Siri and Alexa as AI systems, but, at present, they're really just personal assistants, at least for now. However, tools such as IBM's Watson are here today and are already working behind the scenes in many industries to make better decisions and even to save lives. Watson is being used in healthcare to help doctors make faster diagnoses and to improve patient treatment.

As mentioned, though, data drives this creation and, according to Wikipedia, there are a lot of sources of data.

> The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson als used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used. The IBM team provided Watson with millions of documents, including dictionaries, encyclopedias, and other reference material that it could use to build its knowledge.

The world should expect to see far more uses of AI in the coming years, for better or for worse.

MACHINE LEARNING AND AUTOMATION

One of the original promises of technology was automation. Way back in the 1700s, punch cards were used to control looms. For centuries, people have been looking for ways to leverage technology to make their lives easier, to reduce costs, or even to transform society.

Consider the current race between companies in the autonomous vehicle space. Companies such as Tesla, Uber, and Google get the most attention, but they're far from alone in these efforts. Automobile manufacturers, including Toyota, BMW, Volvo, Nissan, Ford, General Motors, Daimler, Audi, Baidu, Honda, Hyundai, LeEco, and PSA Groupe are all working hard on their own autonomous vehicle efforts. This is a space that will be incredible to watch in the coming decades as it will have a monumental impact on how people travel for work and for leisure.

Just how is data used in an autonomous vehicle setting? Well, take a look at the Figure 3-2. You can see that this is no ordinary vehicle. Autonomous vehicles are outfitted with all kinds of sensors and cameras that feed their output to a server farm that lives in the trunk. All of this data is constantly crunched so that the vehicle knows exactly where it is, where it's going, what obstacles it may encounter, such as road construction, or even surprises such as an animal jumping in front of it.

Photo credit: By Steve Jurvetson - originally posted to Flickr as Hands-free Driving, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=8271620

Figure 3-2. An autonomous vehicle with a full sensor

Over time, as more self-driving vehicles come into production, governments will need to find ways to ensure that these vehicles always have up-to-date information about everything going on around them. These cars will be always-on and always connected to wireless networks. Further, cars will talk to each other in order to ensure the most efficient flow of traffic. The portable data centers residing in the trunks of these vehicles will be in regular communication with a transportation grid as well as other cars in the local vicinity. These cars will share data with each other so everyone stays current and so that each vehicle knows when it needs to take an action, such as stopping at a stop sign. That said, if done correctly, self-driving cars could spell the end of the stop sign and traffic light industries. If cars can communicate with one another, you will see cars weaving in between one another at the exact right speed rather than waiting at red lights.

Some of the above information is happening today and some of it is coming in the future, but all of it relies on a comprehensive data foundation.

BEHAVIOR TRACKING

Have you ever noticed that, after you look at a product or service somewhere on the web that ads for that product or service seem to follow you around? That's not just a coincidence. Read on.

WEB EXPERIENCE

You entire web experience is a data-driven one. Behind the scenes, there are complex data-driven processes taking place. Involving data that is stored on both computer, with your Internet Service Provider, with huge ad networks, and with sites that you visit, this carefully coordinated experience attempts to target advertising at you based on a myriad of factors.

The end result is an ongoing assault of personally targeted calls to action (CTA). Everywhere you go, you're being pushed to buy things that you've already looked at. Why do marketers do this to you?

Because it works… and they have the data to prove it.

Using data-driven audience segmentation tools allows marketers across the spectrum to track behavior and help boost their companies' bottom line. Once a critical mass of audience tracking behavior is captured, these advertisers can begin to use predictive analytics to better target their ads. For them, better targeting results in more clicks, which translates to an increase in revenue.

Remember… big data, in many cases, is all about the bottom line!

NETFLIX

Whether you use it or not, you know all about Netflix. The entertainment company is a big data behemoth. They know everything you watch, when you watch it, what you search for, who in your family is watching, and how much you watch. Netflix tracks everything you see, and everything you click on. The company then throws this data into their big data engine, which spits out carefully tailored recommendations. But, it also impacts the very advertising that you see about the service and the programs on the service.

In 2014, Wired Magazine published an article (https://www.wired.com/insights/2014/03/big-data-lessons-netflix/) about just one tiny aspect of Netflix's big data efforts and it's a fascinating look into just how seriously the company takes these efforts.

The article discusses the color breakdown of the covers that you see for various programs on the service. Netflix isn't just scooping up your data; they're also tracking their own information, such as all of the colors used in each of their ads, and then

comparing that information against user behavior to see if there is any significant difference in audience engagement.

In the Wired article, the author notes that the company makes data visualization tools widely available so that they can gain quick insight into ways to improve the business.

AMAZON RECOMMENDATIONS

Besides being the largest public cloud computing provider on the planet, Amazon is also one of the world's largest retailers. And, like many other large companies, particularly those that live on the web, Amazon grabs as much data about you as possible. Like Netflix with its program, Amazon knows about everything you search and everything you buy. Want proof? Log in to Amazon now and take a look at the recommended products. As you browse the web outside Amazon, take a look at the ads in your sidebar. They will be related to what you were considering buying at Amazon.

On the recommended purchase front, there's more here than just user behavior and activity tracking taking place. Behind the scenes, Amazon has a vast database of product information, including how every product relates to every other product, and much, much more. By coupling that information with your personal activity, Amazon can make granularly targeted recommendations that might just tempt you to click on that Buy button.

SPEECH RECOGNITION AND PERSONAL ASSISTANTS

Although the various personal assistants on the market have varying degrees of popularity and success, these tools, such as Siri, Alexa/Amazon Echo, and Google Home are becoming common fixtures in many households and on mobile devices. As people transition to a voice-driven, on-demand world, such services will continue to grow and will eventually become as indispensable as the smartphone.

Personal assistants rely on comprehensive speech recognition capabilities in order to work. When you ask Alexa, "What's the weather", you reasonably expect her to tell you what the weather is, not what's on the dollar menu at McDonalds.

Here's a quick primer on how speech recognition works:

• You make a request of your system. Your voice is an analog input.
• An analog-to-digital conversion process converts your voice to digital data. In order to be processed, the input must undergo this conversion.
• Behind the scenes, your input is cleaned up as much as possible.
• Your fill request is broken down into tiny chunks that each can be as short as a fraction of a second.
• Each of these segments is then compared to a database of phonemes. A phoneme is the smallest unit of speech and is each phoneme represents a distinct sound present in a language. These are used to differentiate words from one another. There are 44 generally agreed upon phonemes n the English language.
• From there, each phoneme is compared to all of the other adjacent phonemes from your request and the engine attempts to decipher what you said.
• Finally, as long as what you said was intelligible and it was accurately analyzed, your command is executed.

From start to finish, the process outlined above can take just a few seconds, but it relies on mountains of data to work. Over time, these voice-driven systems learn from their mistakes, too. All of the requests to these systems is cataloged, and algorithms are continuously adjusted to improve the accuracy of the output. Sometimes, it might seem that these systems have a mind of their own (and someday, they might!), but, in reality, there is just a lot of hidden complexity that doesn't always go exactly how you expect.

FRAUD DETECTION/RISK MANAGEMENT

If you travel much, you may have experienced what can be a huge hassle for you, but that is a huge money saver for credit card companies: automated fraud detection. Again, these kinds of systems rely on massive quantities of data and revolve around behavior tracking. Over time, as you make purchases, you credit card issuer is tracking all kinds of data around the purchase, from the amount, to the store, to the location, to whether it was in person or online, and much more. Eventually, your issuer will have a profile that reasonably matches your spending habits. Should a transaction be attempted that deviates from that profile, the transaction may be declined.

Recently, though, many issuers have taken to automated systems, again, based on data, to help ease the frustration and embarrassment associated with declined transactions. A transaction may still be declined, but you may receive an immediate text message asking if the transaction was actually valid. You can respond right away and your card will be unlocked and you can attempt the transaction again.

This use of data does two things: 1) it helps credit card and other companies combat fraud; 2) it tries to help limit the inconvenience by immediately contacting you to try to resolve the problem.

How Panoply Helps Organizations Leverage Data

At this point, you may be asking yourself how you can get started supporting your organization's data analysis efforts. The start to these initiatives is to first identify the business need and, once identified, it's time for data engineers to spring into action to prepare infrastructure and database services for use by data scientists. You have choices when it comes to the tools that you will use to achieve the infrastructure part of the equation. You can choose to build something out yourself, or you can turn to the cloud. For many, going directly to Amazon is daunting, and they only fulfill a portion of their analysis needs, and the setup can be painful.

This is where Panoply comes to the rescue! No longer are data engineers left on their own. Panoply has your back!

Let's start with a look at the big picture. Figure 3-3 gives you a look at the overall Panoply architecture.

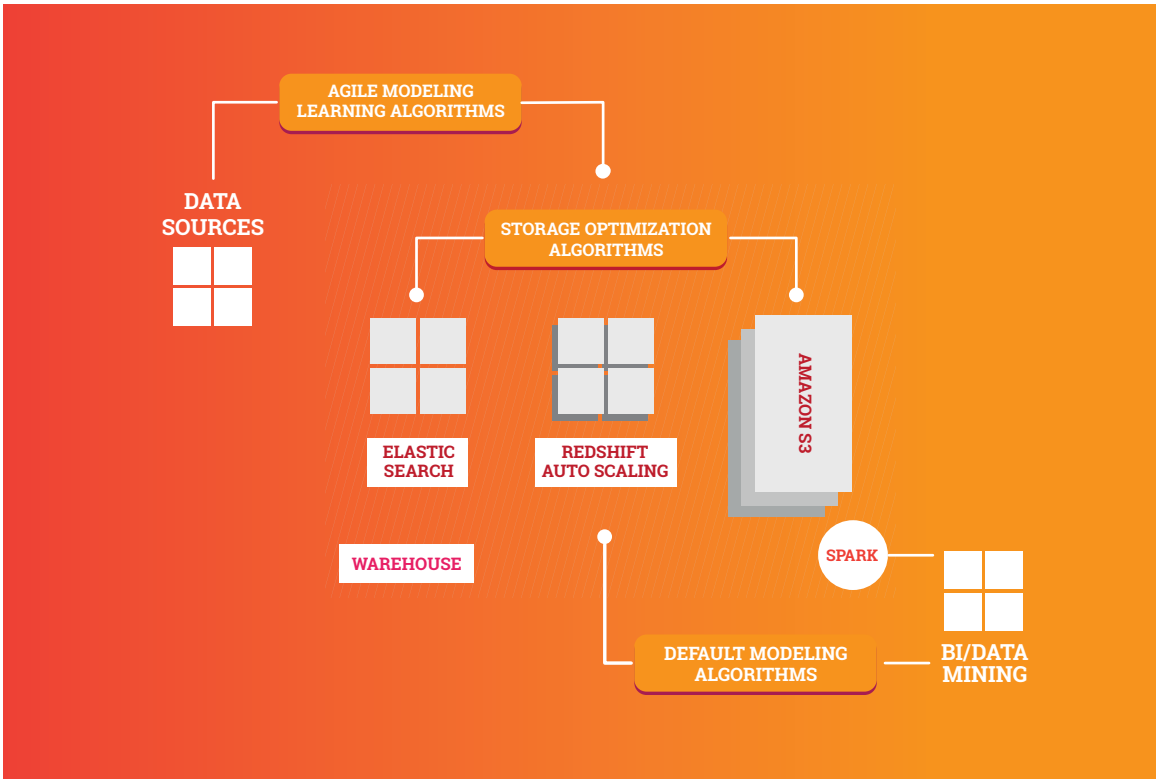


Figure 3-3. The Panoply architecture.

By the end of this paper all of this will make sense.

Now, let's start at the beginning. Panoply.io provides end-to-end data management -as-a-service. Its unique self-optimizing architecture utilizes machine learning and natural language processing (NLP) to model and streamline the data journey from source to analysis, reducing the time from data to value as close as possible to none.

THE PANOPLY ENGINE

The Panoply engine is what makes the whole thing tick. It's comprised of a number of services that are conglomerated to bring to you data goodness.

ETL-LESS DATA INTEGRATION

Panoply.io automatically aggregates data as it streams in, allowing you to analyze everything in seconds – regardless of scale, and without data configuration, schema, or modeling.

Panoply.io offers a collection of pre-defined data source integrations to all of the popular databases and services – open-sourced – and provides an array of SDKs in many of the most common programming languages, so that you can easily tailor the platform to your needs and connect to any data source.

AUTO-GENERATED SCHEMAS

When you insert data into Panoply, the platform scans through the data and discovers the underlying schema and metadata that best describe it – including all columns, data types, and foreign keys. It constructs this schema based on the data, or alters the existing schema in real time (when necessary), thereby eliminating the need to explicitly design database tables and columns.

Panoply.io makes it easy to change data types or add columns – you can simply input different value sets into the platform. If necessary, manual adjustments and customizations can also be made.

REAL TIME TRANSFORMATIONS

Panoply.io uses common transformations automatically, including the identification of data formats like CSV, TSV, JSON, XML, and many log formats – and flattens nested structures like lists and objects into different tables with a one-to-many relationship.

Remodeling and reindexing are also automatic processes, taking place whenever the system detects changes in query patterns. Panoply.io uses statistical analysis to identify the columns and tables that are used most frequently in filters and group-bys, and uses that information to rebuild indexes.

3-TIER STORAGE ARCHITECTURE

Panoply.io has a 3-tier stack of storage systems abstracted away behind a single JDBC end point: AWS S3 is used at the backend, as a massively scalable storage engine for semi-structured data; Redshift is used for most of the data, and especially for structured and frequently accessed tables and rows; and Elasticsearch provides fast access and searches through data and aggregations, and handles the indexing and storage of common daily queries.

STREAMLINED DATA UTILIZATION

Panoply.io delivers a set of pre-integrated, cloud-based analysis tools through a Data Apps framework, which is easily extendable to your own tools and platforms.

Panoply.io exposes a standard JDBC end point with ANSI-SQL support, providing plug-in support to your Tableau, Spark, or R analytics tools. The platform also allows you to write your own SQL code and build apps on top of the data.

SIMPLIFIED USER MANAGEMENT

Panoply.io provides streamlined management of users and permissions, avoiding the cumbersome SQL configuration generally required to manage lists of users, passwords, grants, and denies – and allowing you to send out Invites via an easy-to-use UI.

Panoply.io allows you to specify what permissions users have and which tables they can access, and to view a complete activity log of activities per user – making it easy to pinpoint why changes were made.

ENHANCED PRIVACY & SECURITY

Built on top of AWS, Panoply.io uses the latest security patches and encryption capabilities provided by the underlying platform including permission controls, TLS, and hardware accelerated RSA encryption.

Panoply.io also offers an extra layer of security built to enhance data protection and privacy, that includes columnar encryption, two-step verification, anomaly detection, and handling expiring accounts.

EFFICIENT MONITORING

Panoply.io is a fully managed analytical data platform that provides maximum transparency about everything from uptime and average query time, to low- level details such as the IO throughput of the physical disks.

Panoply.io's monitoring capabilities include an analysis of all queries performed on the data by all users, making it easier to identify bottlenecks, catch unexpected behaviors, and "rewind" a database to any previous point in time.

ADAPTIVE AUTO-SCALING

Panoply.io handles the entire data infrastructure, eliminating the traditional concerns about scale, caching, IOPS, and memory. The platform auto-scales clusters seamlessly to keep up with the organization's needs while reducing server costs.

Panoply.io adapts server configurations over time based on data scale and query patterns – scaling up or scaling out servers, as necessary. Scale changes take place on a regular basis and can occur multiple times throughout the week, optimizing the system's performance.

Summary

Hopefully, your journey through time, from decades ago to today and into the future has provided you with an understanding of where today's data engineering efforts originated. And, you now know that, as a data engineer, your efforts are in constant demand and you need to find ways to deploy architecture upon which data science can take place. With Panoply, you have the opportunity to be the data engineering hero that your organization needs.

Panoply.io provides end-to-end data management-as-a-service. Its unique self-optimizing architecture utilizes machine learning and natural language processing (NLP) to model and streamline the data journey from source to analysis, reducing the time from data to value as close as possible to none.

Panoply.io 188 King St.• San Francisco, CA 94107 • hello@panoply.io