Data Scientists Need Usable Data

By: Katie Horvath, J.D.
VP of Marketing & Communications,
Aunalytics, Inc.

aunalytics

> It is a well-known industry problem that data scientists typically spend at least 80% of their time finding and prepping data instead of analyzing it.

The IBM study originally publishing this statistic dates back to even before most organizations adopted separate best-of-breed applications in functional business units. Typically, today there is not one central data source used by the entire company, but instead there exist multiple data silos throughout an organization due to decentralized purchasing and adoption of applications best suited for a particular use or business function. This means that now data scientists must cobble together data from multiple sources, often having separate "owners" and wrangle IT and the various data owners to extract and get the data to their analytics and then make it usable. This is a complex technical problem and a complex political problem.

"The overwhelming majority of effort a typical data scientist puts forth has to do with creating a clean data set with useful information, all before any of the compelling machine learning or statistical models can be applied." 8 Real Challenges Data Scientists Face, Forbes (2018). In fact, it has been said that data modeling is now 90% gathering and cleaning data and 10% model building. Common Workplace Problems for Data Scientists and How to Address Them, Dataquest (2019). This problem now extends to data analysts, who also spend 90% of their time integrating and harmonizing data to make it usable. *Thomas Goulding, Professor for Master of Professional Studies in Analytics at Northeastern University, Biggest Data Analytics Challenges of 2020, Northeastern University Graduate Program (2019).*

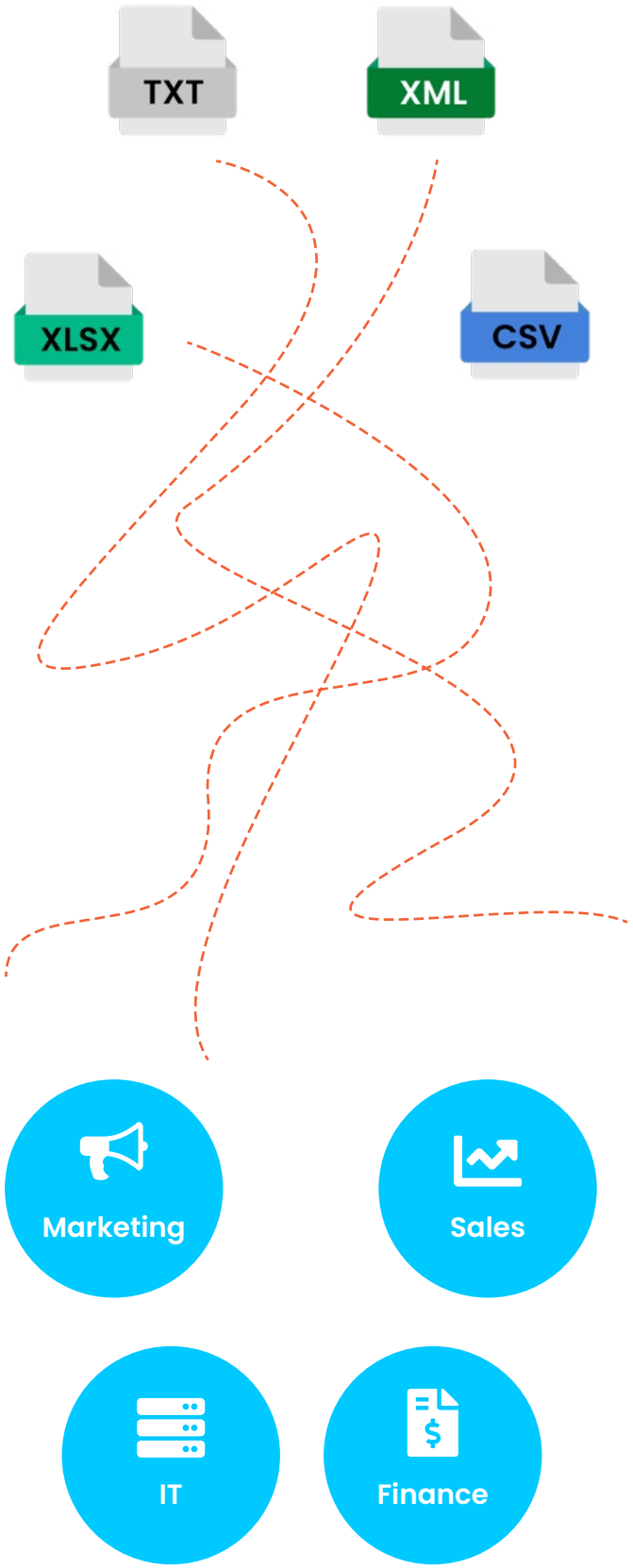
Data munching (turning it from raw data into a consumable form) includes detailed knowledge of database structures (such as Cassandra, MongoDB, SQL), data mapping, data wrangling techniques, fixing issues of inconsistency and quality, and writing code. Then big data considerations come into play and technologies such as Apache Spark are needed to handle processing of large data sets. It falls upon the data scientist, data engineer or IT to glue it all together and often the process is manual. Nearly half of organizations report that the biggest barriers to data analytics success include lack of technical know-how to accomplish this. Other leading factors dooming success include cost, problems with technology and inability to make data usable for end-users. *Most Common Problems Companies Are Facing With Their Big Data Analytics, BI-Survey (2020).*

Getting a pipeline of data to your analytics requires involvement from IT for infrastructure and multiple business unit teams, who are the various "owners" of the data that needs to be combined. Even data merges require approval, as well as the various data hand-offs between teams, and typically different teams have responsibility for different parts of the pipeline. *Sabrina Stanescu, Your First ML Model in Production – Examples and Considerations, Altair (2020).* A common challenge in data science is facilitating cooperation between departments on how data should be collected and interpreted.

> "Predictive models and historical analyses are only as powerful as a team's agreement on the validity of the source data."
>
> *Andrew Seitz, Snowflake Senior Data Analyst quoted in Forbes (2018).*

Imagine having to get all of these stakeholders to agree upon how to get data to you before you can do your job. It's like trying to get all of the world's Master Chefs to agree on one best recipe. Often there is misalignment between engineering and data teams, and from one data owner to another.

In addition to finding and mapping the data, the next challenge is that often the data is not in the right form or format for answering the question posed, exists in multiple different forms and formats that must be normalized for analytics, and just does not match up for comparing apples to apples. It takes significant data scientist time to normalize data into a usable format. It is not surprising that a top data scientist frustration is people assuming that data discovery, collection, cleaning, storing and manipulating is the easy after-thought part of a data scientist's work. *What are the Main Pet Peeves of Data Scientists, Quora (2018).*



In addition to problems with form, all too often the data is of poor quality.

> "Expecting data scientists to take bad data, little data, or no data and turn it into meaningful, actionable predictions is another expectation problem data scientists can face. Managers may have read articles about the power of machine learning and AI and concluded that any data can be fed into an algorithm and turned into valuable business intelligence."
>
> *Common Workplace Problems for Data Scientists and How to Address Them, Dataquest (2019).*

"Data quality and volume are non-negotiable." *Ganes Kasari, CoFounder and Head of Analytics Gramener, What Frustrates Data Scientists in Machine Learning Projects, Towards Data Science (2018).*

Often the data available does not answer the question posed. Either the data scientist is forced to try to fill gaps in the data to create the dataset actually needed to answer the business question or explain to management that the question cannot be answered.

"Your analysis and predictions can only be as good as the data you're working with. Certainly there are statistical techniques that can help you plug gaps in a data set, but there's no magical algorithm that'll predict six months of sales accurately when it's only fed a week of data to learn from." *Id.*

Tracking changes to the data is near impossible with the numerous data owners and sources. Data models blow up due to changes made to data sources that are not communicated to the analysts using the data. For example, if the contents of a field change completely (row A was a zip code and now row A is a phone number and row B is the zip code), the data feeding the algorithm stops making sense. Changes in form or format can cause the data to go out of range (the 5 digit zip code was changed in into a XXXXX-XXXX format). This problem is made larger by the multiple data owners and by upstream data users not appreciating how a seemingly small change can greatly impact downstream data users. Typically, data analytics fail where data governance over data sources is lacking.

Problems with the underlying data lead to distrust with analytics results. In a recent Massey University study, nearly two thirds of surveyed business management said that they had no confidence or trust in big data, preferring to rely on intuition and personal experience to make decisions, instead of analytics results. *Study Shows Many Senior Managers Distrust Big Data, Massey University (2019).* This means that despite spending a fortune on data analytics solutions, decision-making is not following analytics results.

After all, garbage in, garbage out. If the underlying data is erroneous, the analytics results will not be based on the actual facts at hand. And when AI and ML come into play and algorithms are being trained upon data errors, garbage in leads to exponential garbage out. Lack of trust causes data scientists to have problems convincing management of the value of an analytics project and spend time having to fight for and defend each analytics project instead of getting to work on one. Again, it goes back to the underlying data and without trust in the data, the analytics is of little value.

At a high level, data trust comes from data accuracy and consistency across an organization. If you have different competing addresses for a customer, which is right? Unless all functional business units within an organization are operating off of the same set of facts, time is wasted in tracking down accurate information to be able to answer simple business questions and feed complex analytics. Trust is not a one-time question. To be valuable, analytics cannot be based upon static data. Rather, analytics needs to be ready for changing data. To answer critical decisions in a timely manner, trusted data needs to be updated quickly, and data sources monitored for instability, changes in data base fields, and there must be governance over changes made to the data.

Trust also stems from time spent on wrangling and assessing context and metadata. Sometimes data analysts feel pressured by management to cut corners on wrangling due to project time and cost. "Many data scientists overlook the importance of data wrangling and assessing the context of the data before opening their model toolbox. Hence they miss seeing the risk when clients ask for cutting out 'unnecessary analysis' from the critical path, in order to save precious project time." *Ganes Kasari, What Frustrates Data Scientists in Machine Learning Projects, Towards Data Science (2018).* "Data exploration and analysis are mandatory pre-steps to machine learning and all other advanced techniques. Without getting a feel for the data, discovering outliers or spotting underlying patterns, models do nothing but shoot in the dark." *Id.*

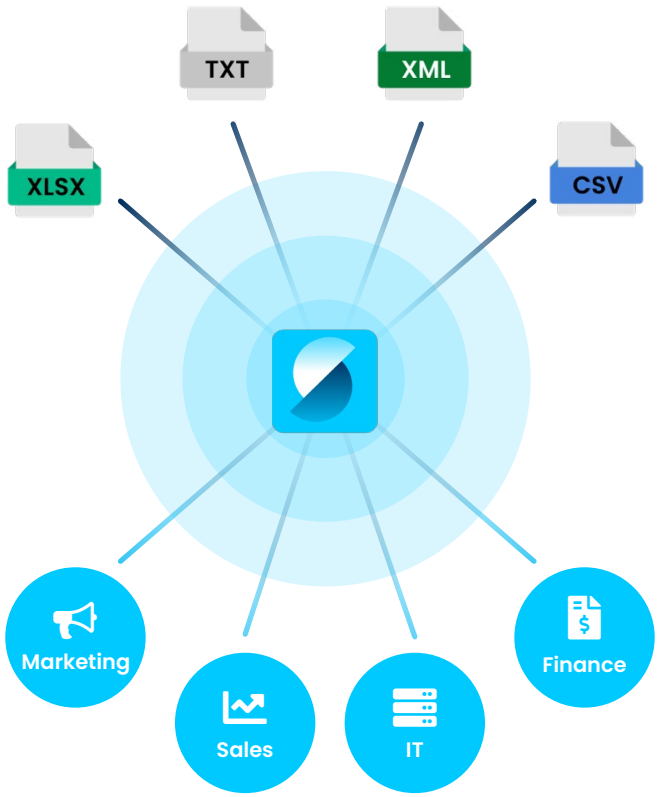
But building data pipelines is not the best use of a data scientist's time. For visualizing analytics results, a data analyst is not expected to build a new dashboard application. Instead, Tableau, Power BI and other out of the box solutions are used. So why require a data analyst to build data pipelines instead of using an out of the box solution that can build, integrate and wrangle pipelines of data in a matter of minutes? Data scientists instead need to spend their highly compensated time developing models, examining statistical results, comparing models, checking interpretations, and iterating on the models. This is particularly important given that there exists a labor shortage of these highly skilled workers in both North America and Europe and adding more FTEs into the cost of analytics projects makes them harder for management to justify. Without investment in automation and data democratization, the rate at which you can execute on data analytics use cases — and realize the business value — is directly proportionate to the number of data engineers, data scientists, and data analysts you hire. This scalability issue dramatically increases the cost of data analytics. *The Problems Facing Data Analytics Customers Today, Data Driven Investor (2019).*

It makes sense to have a centralized golden record of company data that is kept accurate and analytics-ready in real-time, to feed all business units in the organization.

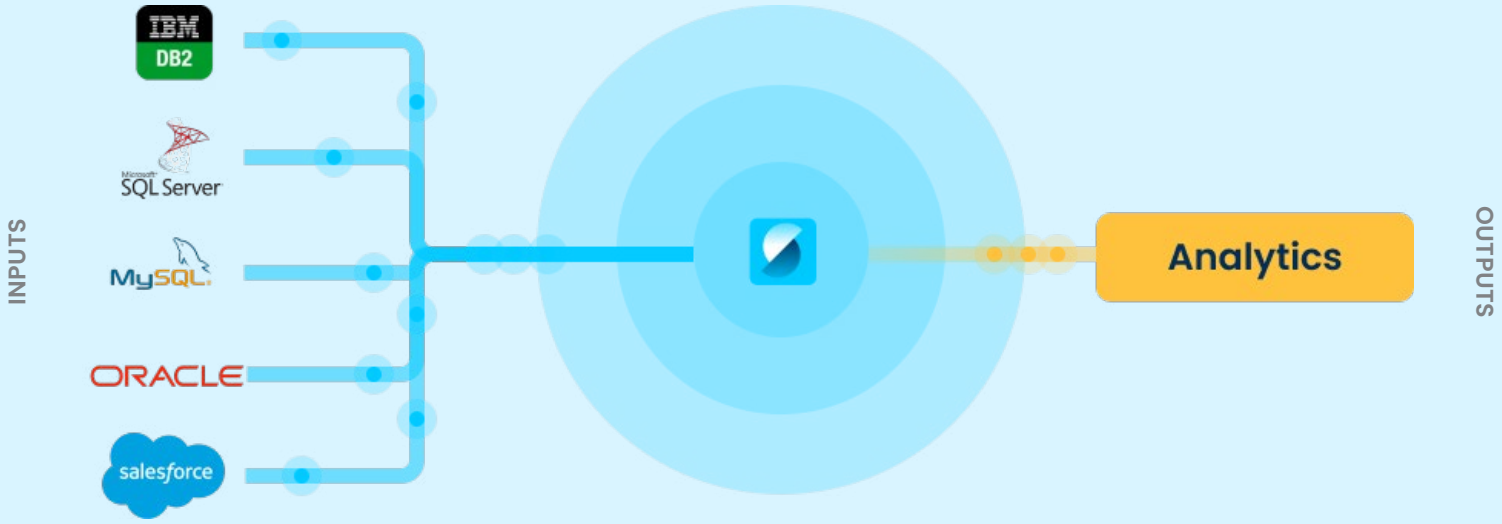
> "Data democratisation means that everybody has access to data and there are no gatekeepers that create a bottleneck at the gateway to the data. It requires that we accompany the access with an easy way for people to understand the data so that they can use it to expedite decision-making and uncover opportunities for an organization. The goal is to have anybody use data at any time to make decisions with no barriers to access or understanding."
>
> *Bernard Marr, Big Data in Practice (2015).*

Thus, a data management solution must connect to all company data sources and automate data normalization for consistency and accuracy, yet not require highly skilled technical resources to use and maintain the system. A self-service no code data management platform that automatically profiles data, detects data quality issues, builds golden records of accurate data from multiple sources and automatically tracks lineage and governance, eliminates the technical and the political barriers faced by data analysts and allows analysts to get straight to finding insights in data that is trusted to be accurate and up to date.

It saves time, money and angst to automate data pipeline building and wrangling, for analytics-ready data to be delivered to the analysts. It saves time and money for a real-time stream of integrated, munched and cleansed data to be automatically created and fed into analytics to give current and trusted data available for immediate decision-making. It saves time and money to have automated data profiling to give analysts insights about metadata, outliers and ranges and automatically detect data quality issues. It saves time and money to have built-in data governance and lineage so that a single piece of data can be tracked to its source. It saves time and money when changes made to source data are automatically detected and analytics are updated without blowing up algorithms.



> "Using a data platform with built-in data integration and cleansing to automatically create analytics-ready pipelines of business information allows data scientists to concentrate on creating analytical results. This enables us to rapidly build insights based upon trusted data and deliver those insights to clients in a fraction of the time that it would take if we had to manually wrangle the data."
>
> *David Cieslak, Ph.D., Chief Data Scientist, Aunalytics*



And taking it to a new level, a cloud-native data management system having pre-built API plug-ins to connect data from multiple cloud environments and on premises sources effortlessly in a matter of minutes, further saves time and money on getting clean data to analytics. Auto-mapping of data sources to a golden record, saves months of services hours typically spent on mapping projects.

And importantly, it gets you your data now and not months down the road. Truly, a plug and play system, agnostic to data source type and format that allows users with no technical knowledge to create golden records by a drag and drop interface, that gives you immediate results and keeps the analytics ready data updated in real-time is the next generation of data management technology.

About Aunalytics

Aunalytics is a data platform company delivering answers for your business. Aunalytics provides Insights-as-a-Service to answer enterprise and midsized companies' most important IT and business questions. The Aunalytics® cloud-native data platform is built for universal data access, advanced analytics and AI while unifying disparate data silos into a single golden record of accurate, actionable business information. Its Daybreak™ industry intelligent data mart combined with the power of the Aunalytics data platform provides industry-specific data models with built-in queries and AI to ensure access to timely, accurate data and answers to critical business and IT questions. Through its side-by-side digital transformation model, Aunalytics provides on-demand scalable access to technology, data science, and AI experts to seamlessly transform customers businesses. To learn more contact us at +1 855-799-DATA or visit Aunalytics at https://www.aunalytics.com or on Twitter and LinkedIn.

About the Author

Katie Horvath led a data management company as President & CEO, recognized as the only woman CEO of a big data company in North America, until 2021 when Aunalytics acquired her company. Katie has been recognized at U.S. Congress for innovative business models. She has a passion for leading and scaling tech companies, and making complex technology understandable and usable for non-technical business users. Katie is an IP lawyer and an engineer.

aunalytics