Four Data Engineering Fundamentals All Data Scientists Must Know

DATA ENGINEERING DATA SCIENCE INTERMEDIATE PROFILE BUILDING

This article was published as a part of the Data Science Blogathon

Introduction

Data Science is a team sport, we have members adding value across the analytics/data science lifecycle so that it can drive the transformation by solving challenging business problems.

We have multiple team members in a data science team: **data engineers who create the foundation of all data** that is consumed by analysts to explore and do descriptive analytics further advanced ML models created by Data scientists – visualized by BI engineers & deployed by ML engineers. All of them must work in tandem to successfully drive an organization's data science program.

A typical stakeholder map for the data science team is mentioned below:

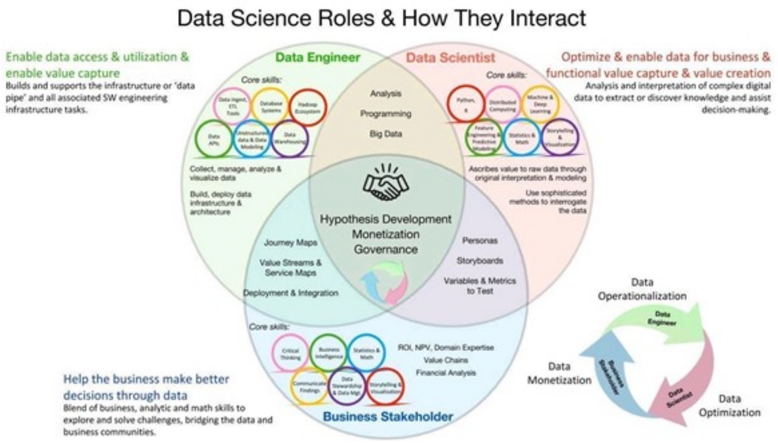
Data Science Roles & How They Interact
Enable data access & utilization & enable value capture
Builds and supports the infrastructure or 'data pipe' and all associated SW engineering infrastructure tasks.
Data Engineer
Core skills:
Data Ingest, ETL Tools
Database Systems
Hadoop Ecosystem
Data APIs
Unstructured data & Data Modeling
Data Warehousing
Collect, manage, analyze & visualize data
Build, deploy data infrastructure & architecture
Data Scientist
Optimize & enable data for business & functional value capture & value creation
Analysis and interpretation of complex digital data to extract or discover knowledge and assist decision-making.
Core skills:
Python, R
Distributed Computing
Machine & Deep Learning
Feature Engineering & Predictive Modeling
Statistics & Math
Storytelling & Visualization
Ascribes value to raw data through original interpretation & modeling
Use sophisticated methods to interrogate the data
Analysis
Programming
Big Data
Hypothesis Development
Monetization
Governance
Journey Maps
Value Streams & Service Maps
Deployment & Integration
Personas
Storyboards
Variables & Metrics to Test
Core skills:
Critical Thinking
Business Intelligence
Statistics & Math
Communicate Findings
Data Stewardship & Data Mgt
Storytelling & Visualization
ROI, NPV, Domain Expertise
Value Chains
Financial Analysis
Business Stakeholder
Help the business make better decisions through data
Blend of business, analytic and math skills to explore and solve challenges, bridging the data and business communities.
Data Operationalization
Data Engineer
Data Monetization
Business Stakeholder
Data Scientist
Data Optimization

Image 1

Contents:

Now the question – if we have champion data engineers in the team why do data scientists need to know those data engineering/data management concepts?

1. They are consumers of the data hence **to create robust analytics solutions with that data** – knowing when and how the data is collected, stored and prepared helps them get the right ways and tools to pull data, derive insights, and design models
2. Data science teams might need **to interact regularly with data engineering teams** to get new data, share additional data information for derived tables – knowing these concepts enables to have a more efficient conversation
3. There has been a higher emphasis on using data with consent and per regulations. Data science teams should be (they already are) closely involved with data regulations so having this knowledge would help **to stay compliant and reduce the risk of data regulations**

In a nutshell, data science teams need to play their role of being able to efficiently derive the best value from (big) data without compromising on data regulations, and knowing data engineering concepts helps them do so better.

With that context, let's jump straight into a view at the concepts from a data scientists' lens!

Data Warehouse and Data Lakes

What data scientists might not know:

While learning to design dashboards and create models, data scientists are more familiar with it is based on data stored in data warehouses and sourced from data lakes. Data scientists might not know what the best techniques are to query data from the warehouse and what is the best way to look at that data holistically.

Key fundamentals[i]

- The data warehouse is the centralized source of truth database created from multiple sources (each department might still have its own warehouse) (eg. Financial services industry data like credit card transactions)
- Usually having a denormalized structure (for faster queries) and each table has been prepared and structured for a potential business case
- Data Lakes are a step before data warehouses where raw data (including unstructured) is stored, all of the data is kept even if its purpose might not have been defined yet. (eg. Clinicians notes in healthcare)

How it helps data scientists[ii]

- An ML model/analytics solution is as good as its data so it becomes imperative for data scientists to know the origins of data
- In most data science projects 80% of the time is spent in data wrangling so knowledge of data warehouse and then able to understand/ create/ request for analytics-ready data sets/ datamarts can help increase efficiency and reduce project timelines
- Data lakes can help data scientists in discovery exercises to identify data for use cases

Data ETL (Extract Transform Load)/ Pipelines

What data scientists might not know:

Data that is collected and the ones presented for analysis has often a lot of preprocessing and transfer steps involved before it lands up in the data warehouse or analysis file. Most of the data scientists while learning ML / AI might have used already prepared data which eliminates the need for but in actual ML design in an industry often the data scientist has to prepare and modify data per the use case – they definitely need to know what was the data that was collected and how it ended up in a specific field (for example does Null gender mean that use did not want to share it or does it mean the data was unavailable or both – the data engineering team would have those answers)

Key fundamentals[iii]

- ETL = "extract, transform, and load," are the data engineering steps that are required in data preparation whether to store it in a warehouse or to use it for an ML model/analytics use case
- It involves getting data from a source (eg. Adobe analytics on the website that is stored in Adobe cloud) to preparing a data feed from it and then transforming it into a format that is relevant for the business (integration with the organization's unique customer id, for example, changing currency to $ form local currency) and then loading it into one or multiple tables in data warehouse/ lake. Sometimes the transformation is done after loading that data and it is called ELT.
- A data pipeline is the series of connections and steps through which data moves from one location to another
- The data feed is a block of data that is ingested into the data warehouse periodically through ETL processes

How it helps data scientists

- ML Models/analytics solutions are not just made for onetime but need to be constantly updated and refreshed – for that ML and data pipelines need to
- Data ETL concepts can be applied in ML pre-processing to make production-ready code and workflows that can be used during ML implementation
- Knowledge of ETL processes can help in data lineage understanding and right interpretation of data (eg. Knowledge of 'Age' data was collected at the point of sale manually or automated and mapping applied for ageàage bands before storing can help better design ML models)

Image 2

Data Governance and Quality

What data scientists might not know:

Data is the foundation of all analytics solutions, if even a portion of the dataset is altered it completely damages any downstream models etc. created, often there are no checks to logically check data consistency for a particular context (eg. If suddenly revenue per customer increases from $100 to $800 without any change in the business environment, then it would lead to wrong ML scores and incorrect dashboards). Therefore, a data science team must work closely with the data governance and engineering team to set checks along all critical paths to ensure all models and analytics consistently get the right data.

Key fundamentals[iv]

- Data governance is a broader term that is used to define how organizations manage data objectives, scope, ownership, privacy, and security including standardized process and data
- Data quality is a subset of data governance focusing on continuous monitoring of data for completeness, consistency, and plans to handle data irregularities

- Eg – if an organization has to ingest social media data then data governance would conduct all assessment and planning under data governance and then assess the data received using data quality.

How it helps data scientists

- Data quality helps create robust analytics solutions and keep the reputation and confidence of data science teams
- It prevents re-work and wrong business decisions if proactively identified and jointly solved by IT, data science, and business teams
- It is like model output monitoring but in this case, the input data to a data warehouse is closely monitored to alarm for any irregularities

Image 3

Data Regulations and Ethics

What data scientists might not know:

The data being used might be constrained by legalities and even the ML models created might have bias and used the data in an unintended way sometimes not complying with ethical standards. Any legal implications or brand image incidents might be something that was driven by work that a data science team has done. Since the data science team took lead in handling data and analytics solutions of that data, they are responsible for its impact. Surprisingly many analytics teams do not know this and are not prepared for it. User consent might not have been collected for the use case which the DS team has used for.

Key fundamentals[v]

- Data regulations refer to rules governing the collection, disclosures, storage, usage, and then clearance of data at end of its usage cycle (eg GDPR, CCPA)
- Data Ethics refers to ethical usage, transparency, non-biasness, and righteous usage of data (eg. Not using societal strata data to reject customer loans even if that certain stratum might have bad repayment history)

How it helps data scientists

- Prevents the legal, brand, and reputational risks of using data in the right way
- Helps develop customer-friendly models that can serve as examples across the organization
- Better manage access to sensitive data sharing across the teams to avoid data sharing in wrong hands thereby helping in a better data governance strategy design

Closing Thoughts

Analytics Stack: Brings it all together – It combines all the elements (4 mentioned here) to a single entity that the analytics team consumes to produce results. Typically, it would look like below with some variations.

Image 4

A data science team must focus on these four factors to build a resilient and stable practice and keep adding value to the business with high quality.

References

[i] https://www.talend.com/resources/data-lake-vs-data-warehouse/

[ii] https://towardsdatascience.com/data-warehouse-68ec63eecf78

[iii] https://www.snowflake.com/guides/etl-pipeline

[iv] https://www.collibra.com/blog/data-quality-vs-data-governance

[v] https://www.datascience-pm.com/10-data-science-ethics-questions/

Image Sources-

1. Image 1: https://medium.com/co-learning-lounge/job-roles-in-data-science-10e790ea21b5
2. Image 2: https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5
3. Image 3: https://www.edq.com/blog/data-quality-vs-data-governance/
4. Image 4: https://www.tellius.com/the-modern-data-analytics-stack/

Article by **Ashwini Kumar** | Data Science Lead & Crusader | Linkedin

The media shown in this article are not owned by Analytics Vidhya and are used at the Author's discretion.

Article Url - https://www.analyticsvidhya.com/blog/2021/09/four-data-engineering-fundamentals-all-data-scientists-must-know/

ASHWINI KUMAR