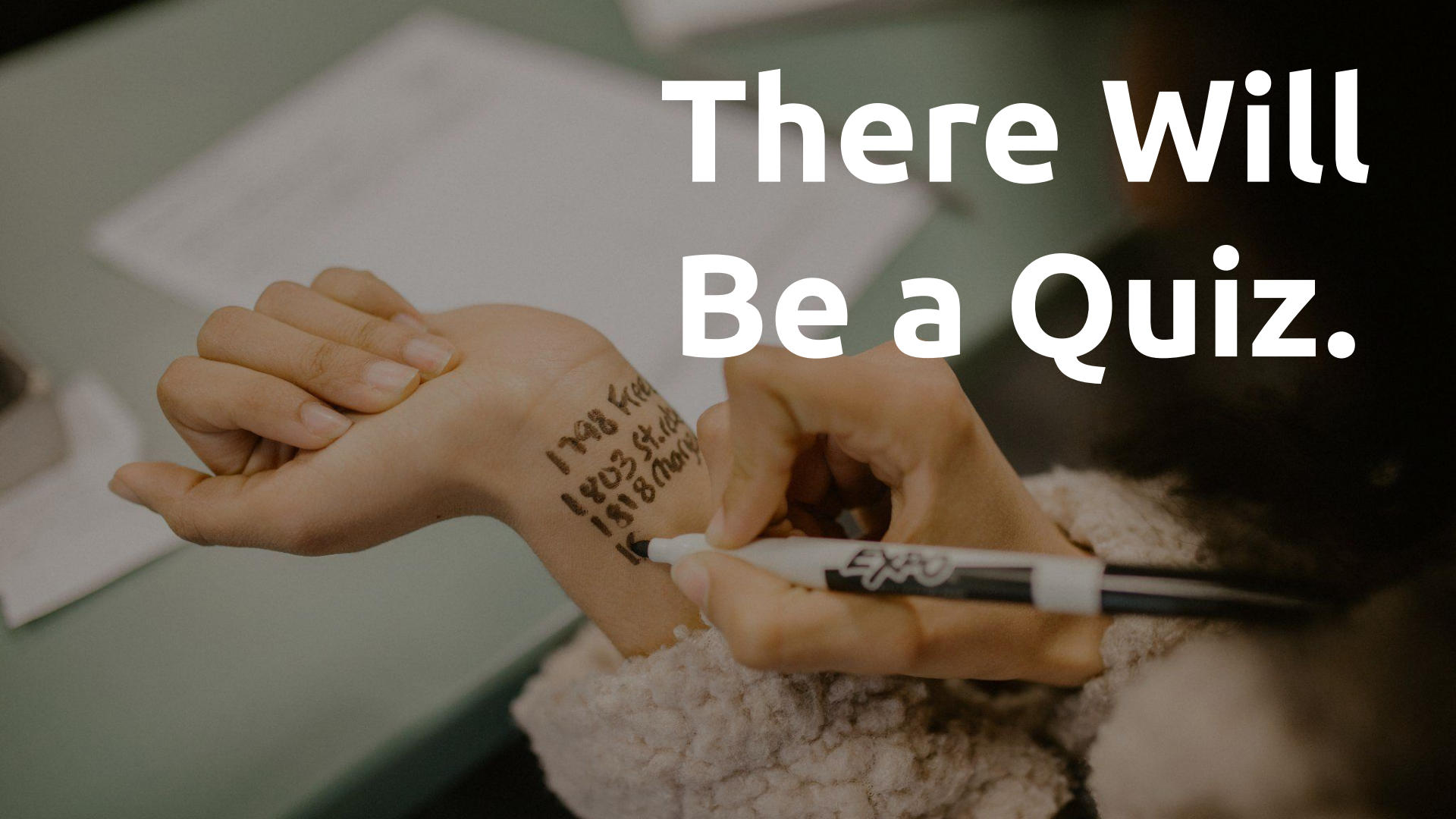
Inciter

Data Engineers Are from Mars and Everyone Else is from Venus

Taj Carson, CEO, Inciter

Data. Driven.


There Will
Be a Quiz.

It can be hard to communicate with, and work with, data engineers to get the results you need.

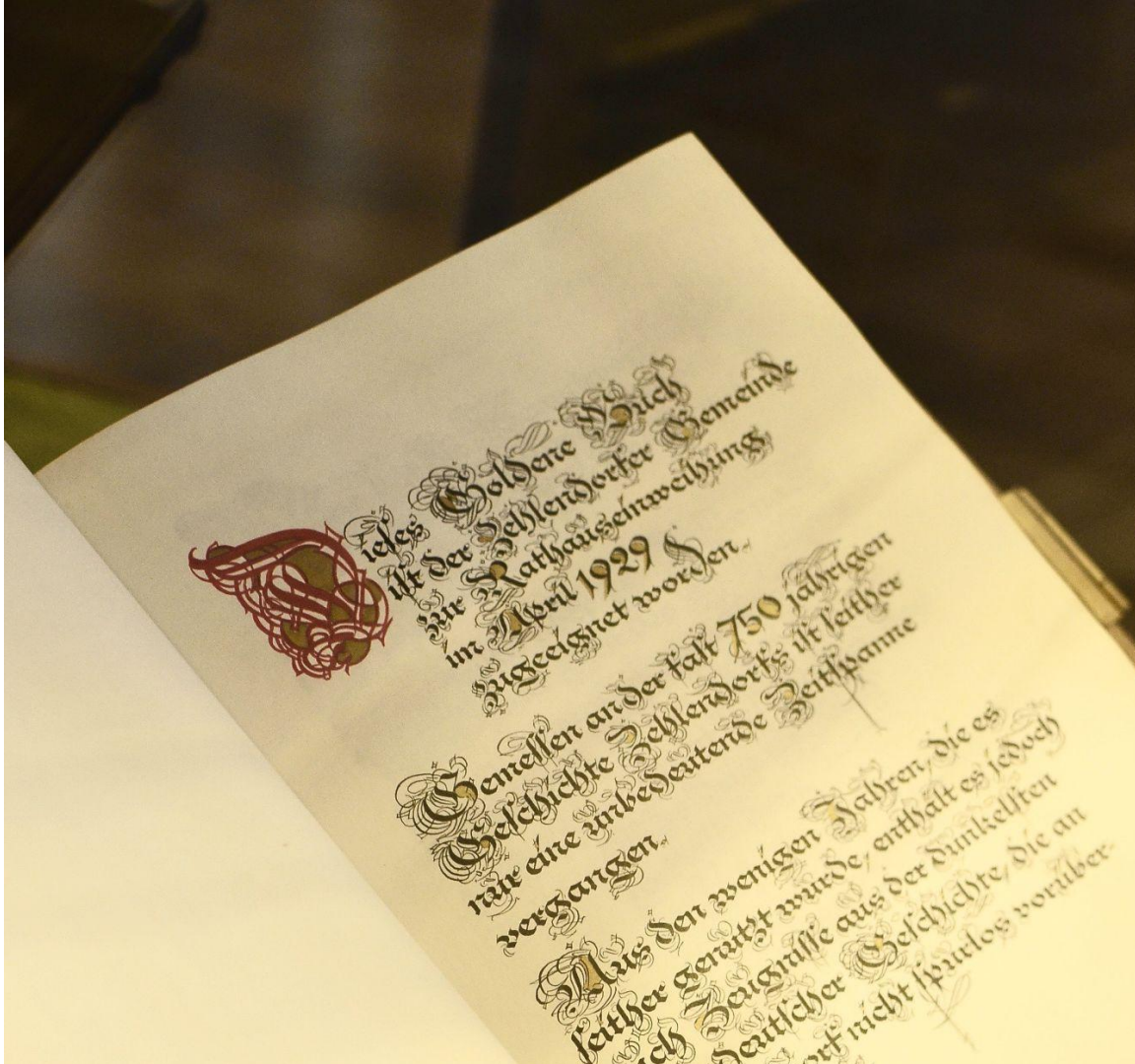
Reasons why you might want to work with a data engineer

- You have data in many different places and want to bring it together into one place
- You've got data in different systems and you need to push or pull data from one system to another
- You've got some labor intensive processes with large and/or repeatable data that you'd like to automate

Data Engineers and Analysts Have Different Mental Models

- You might get more than you wanted…
- Or less.

What if you could translate?




I Need Two Volunteers!
It’s this or another Zoom call…

You are an advocacy organization that focuses on improving access to low income housing. Your have a donor database in **Salesforce**, this is where you store your planned giving data. You also use **MailChimp** for your digital marketing. Finally, you host a conference every year and use **Cvent** to store registration information.

You want to know whether people who respond to your digital marketing campaign become planned giving donors. You also want to know whether conference attendees become donors.

Who is Your Data Engineer?

Where did they work before?

Engineer, analyst scientist?

Engineer or Analyst?

Engineer

- More likely to focus on pipelines, databases, and warehouses (or lakes)
- More likely to manage your Amazon or Azure warehouse

Analyst

- More likely to crunch and deliver the data for reports and analysis
- More likely to work with BI tools

How is the nonprofit sector different for a Data Engineer?

- Smaller data sets - lots more Excel/CSV
- Messier data sets
- Less savvy end users
- Often less automation
- WAY fewer resources

What Tools Do They Use?

- AWS
- Azure
- Google
- Python/R
- IPaaS Tools
- Snowflake

Start With Goals
To make sure you get there

Understand the Constraints
To get the right solution

Ask and Expect Questions
To be clear before you start

Be Specific
To save time

Bring Your Experience
To help them understand

Allow Time
To get the best result

Start with Goals

- Do you need to explore data?
- Create reports?
- Build something to push/pull data?

Start With Goals
To make sure you get there

Understand the Constraints
To get the right solution

Ask and Expect Questions
To be clear before you start

Be Specific
To save time

Bring Your Experience
To help them understand

Allow Time
To get the best result

Project Constraints

- Does data need to be manipulated manually or automated?
- Is this repeatable?
- Output to tables or BI tool?
- Direct connection or uploads?

Start With Goals
To make sure you get there

Understand the Constraints
To get the right solution

Ask and Expect Questions
To be clear before you start

Be Specific
To save time

Bring Your Experience
To help them understand

Allow Time
To get the best result

Ask and Expect Questions

- What exactly are we looking for?
- Where did this data come from?
- What kind of data is it?
- What are you hoping to understand when we are done?
- What kind of analysis do you need?
- What are the deliverables?

Start With Goals
To make sure you get there

Understand the Constraints
To get the right solution

Ask and Expect Questions
To be clear before you start

Be Specific
To save time

Bring Your Experience
To help them understand

Allow Time
To get the best result

Be Specific

- What data sets?
- What fields?
- What are the integrations?
- Where does the output go?
- What calculations (transformations) are needed?

Start With Goals
To make sure you get there

Understand the Constraints
To get the right solution

Ask and Expect Questions
To be clear before you start

Be Specific
To save time

Bring Your Experience
To help them understand

Allow Time
To get the best result

Bring Your Experience to the Table

- **DON’T assume that you won’t understand what they tell you. You are a subject matter expert in your own right just as they are. Communication is key. Confidence is also needed.**

Start With Goals
To make sure you get there

Understand the Constraints
To get the right solution

Ask and Expect Questions
To be clear before you start

Be Specific
To save time

Bring Your Experience
To help them understand

Allow Time
To get the best result

Give it Time.

- DO make sure to allocate enough time for conversations, this takes TIME
  - Time to get to know the person, to have those conversations, to clarify
  - Spend the time up front, and it will save you time later

A Bit of a Martian Dictionary

- Data Lake
- Data Warehouse
- Data Pipeline
- API
- Machine Learning
- Artificial Intelligence

Data Warehouse, Data Lake, Database…..

These days, most of us are using a variety of vendor services, hosted in the cloud. Each one does it's job well, but causes problems when it comes to storing and accessing data:

- Silos of unconnected data that do not talk to each other
- No single version of the truth that combines data in a standard way
- Limited ability to read large amounts of data for analytics; these transactional systems are good at writing data but not good at reading lots of data

Data Lake

A data lake is a vast pool of raw data, the purpose for which is not yet defined.

- It is a type of database, usually with data from many sources, all in one place. If you have one data source, you probably don’t need a data lake.

Data Warehouse

The Data Warehouse sits on top of the data lake. Data warehouses are largely for storing structured data. Data lakes are more of a giant inbox, where you eventually pull out the relevant data to make it useful.

DATA LAKE vs DATA WAREHOUSE

Data

unstructured

Users

Data Scientists,
Data Analysts

Use cases

Stream Processing,
Machine Learning,
Real time analysis

Raw

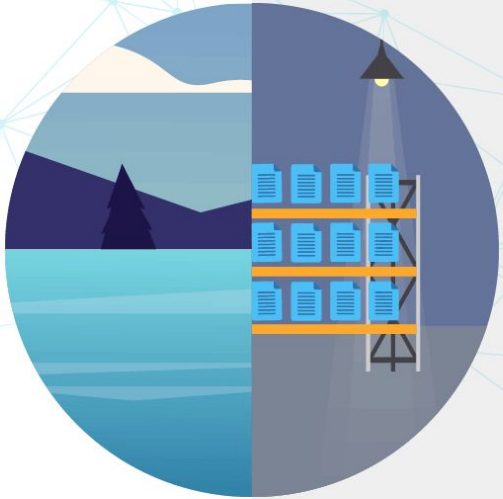
Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI



Data

Structured

Users

Business Analysts

Use cases

Batch Processing,
BI, Reporting

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc

https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-ar

Data Pipeline

A data pipeline is the pathway from a raw data source to the end data consumers. Think of the very real water pipelines that deliver water to your kitchen sink. The water is first pumped out of a source, an aquifer, river, lake, etc, and purified at a water treatment facility. It's then pumped into a storage somewhere like a water tower or a reservoir before the water flows to your home and out your faucet. Data pipelines work the same way.

- We collect data from a source, like a survey or a mailing list. We clean the data with code/programming and "pump" the data into storage, a data lake, database or data warehouse. Finally, the cleaned data flow into reports for the end user

APIs…pretty cool, but not a miracle.

- An Application Program Interface (API) is a software package with the primary purpose to allow a third party to access to the data behind the application
- An API defines the protocols that a programmer can use to request data from an application. It also defines how to send data back to the application

APIs…pretty cool, but not a miracle.

- APIs are helpful because they allow programmers to automate how data is sent to and from an application, and can be used to allow applications to talk to each other

APIs…pretty cool, but not a miracle.

- A stable, robust API supported by someone who knows how to manage it can be a thing of beauty, data flowing back and forth getting updated “automatically”

APIs…pretty cool, but not a miracle.

- But they are technically complex, and if you don’t have a stable API or a person to maintain it, sometimes a spreadsheet works just as well

Artificial Intelligence

- Artificial intelligence enables a machine to simulate human behavior. It is essentially a system that seems smart. Artificial intelligence is where a machine can imitate human-like behavior
- These behaviors include problem-solving, learning, and planning, for example, which are achieved through analyzing data and identifying patterns within it in order to replicate those behaviors

Machine Learning and Artificial Intelligence

- Machine learning is a **type** of artificial intelligence. Artificial intelligence is the overall appearance of being smart, machine learning is machines taking in data and learning things about the world that would be difficult for humans to do.
  - Usually best when you have large quantities of data.

But Beware of Bias…

- This can be bias that reflects our bias
- Or bias based on repeating identified patterns
- Or bias because the algorithm is just wrong.
  - Google Photos and Monkeys/Apes
    - Also Facebook
  - Microsoft's Tay ChatBot
  - AI for Screening Applicants

Two Other Things to Consider

- Tasking
- Documentation


I Need Three Contestants!
Still better than a Zoom call.

$?

Why can it be hard to collaborate with a data engineer sometimes?

A: They have their own language and mental models.

B: They don't like to talk to people.

C: They are, literally, from Mars.

D: They often got their start in the private sector.

$?

Why can it be hard to collaborate with a data engineer sometimes?

A: They have their own language and mental models.

B: They don't like to talk to people.

C: They are, literally, from Mars.

D: They often got their start in the private sector.

Mars and Venus…

Like any expert (including you), data engineers have their own language and mental models. Every field has its own jargon, and data engineering is no different.

They also work in what I like to call invisible spaces, in places that people without their expertise cannot see into. So they often need to use metaphors and analogies to make it clear to others (like lakes, warehouses, and pipelines).

$?

When should you use a data warehouse?

A: When you need to clean your data.

B: When you have a large enough budget.

C: When you have multiple sources of data in different formats.

D: When your data engineer says so.

$?

When should you use a data warehouse?

A: When you need to clean your data.

B: When you have a large enough budget.

C: When you have multiple sources of data in different formats.

D: When your data engineer says so.

Data Warehouses are Powerful

They also require a great deal more expertise to operate and maintain. Consider their benefits and choose carefully.

$?

What's the most important thing to do, to have a successful relationship with your data wizard?

A: Communicate (and ask questions!)

B: Go to data engineering boot camp.

C: Make sure to learn all the technical terms you can.

D: Bring them coffee.

$?

What's the most important thing to do, to have a successful relationship with your data wizard?

A: Communicate (and ask questions!)

B: Go to data engineering boot camp.

C: Make sure to learn all the technical terms you can.

D: Bring them coffee.

It’s really not that different…

Communication, good listening skills, and the power of a well-placed question will serve you here as anywhere else.

Imagine…

Imagine having a data engineer who can create pipelines, automate data cleaning, and integrate your data without you lifting a finger…and you can confidently ask for what you need, clarifying the results you want, and ask and answer questions

You will **BOTH** be more effective, efficient, and successful. And not frustrated!

Inciter

Thank you

Taj Carson, CEO
taj@inciter.io
www.inciter.io

Data. Driven.