



Level 5 Data Engineer Module 6 Topic 4

Data Ingestion Risks

```
31 self.file = None
32 self.fingerprints = set()
33 self.logdups = True
34 self.debug = debug
35 self.logger = logging.getLogger(__name__)
36 if path:
37     self.file = open(os.path.join(path, 'requests.txt'),
38                     'a')
39     self.fingerprints.update([x.request for x in self.requests])
40
41
42 @classmethod
43 def from_settings(cls, settings):
44     debug = settings.getboolean('SUPERMAN_DEBUG')
45     return cls(job_dir(settings), debug)
46
47 def request_seen(self, request):
48     fp = self.request_fingerprint(request)
49     if fp in self.fingerprints:
50         return True
51     self.fingerprints.add(fp)
52     if self.file:
53         self.file.write(fp + os.linesep)
54
55 def request_fingerprint(self, request):
56     return request_fingerprint(request)
```

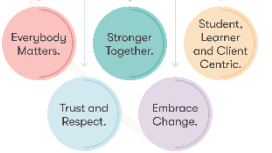
L5 Data Engineer Higher Apprenticeship
Module 6 / 12 (“Data Collection and Ingestion pt. 1”)
Topic 4 / 4

Ice breaker: Discussion

A bit of fun to start...

1. If you could set up an SLA (Service Level Agreement) for any aspect of your daily life, what would it be and why?
2. Imagine you had to manage data ingestion for a personal project. What kind of data would you collect, and how would you ensure its quality?
3. What do you think is the biggest challenge in managing data ingestion risks, and how would you address it?

Building Careers
Through Education



**Submit your responses to the
chat or turn on your
microphone**



Case study

Facebook's Data Ingestion and Risk Management

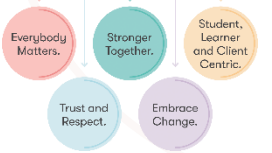
Facebook uses extensive data ingestion processes to collect data from various sources, including:

- User interactions
- Posts
- External websites
- Implemented SLAs to manage and mitigate data ingestion risks

Impact:

- Ensures timely data collection
- Maintains data quality
- Addresses security vulnerabilities
- Manages data risks effectively
- Ensures compliance
- Enhances user trust

Building Careers
Through Education



Knowledge check poll

What is a primary benefit of setting up SLAs around data collection and ingestion?

- A. It simplifies the data ingestion process.
- B. It ensures timely data collection and addresses potential risks.
- C. It reduces the need for data validation.
- D. It eliminates the need for data quality checks.

Feedback: B – It ensures timely data collection and addresses potential risks.

Building Careers
Through Education



Submit your responses to
the chat!

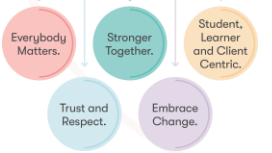


Session aim and objectives

Completion of this topic supports the following outcomes:

1. Explain the benefits of setting up SLAs around collecting and ingesting data
2. Explain how you would manage and report on data risks in a business environment
3. Apply strategies for collecting and ingesting PII and sensitive data
4. Evidence data quality improvement practices in practical scenarios
5. Compress ingested data in Python to reduce sustainability risks and support net-zero goals

Building Careers
Through Education



E-learning Recap

```
31
32 self.file = None
33 self.fingerprints = set()
34 self.logdupes = True
35 self.debug = debug
36 self.logger = logging.getLogger(__name__)
37 if path:
38     self.file = open(os.path.join(path, 'requests.log'),
39                     'a')
40     self.fingerprints.update(ex.request() for ex in self.files)
41
42 @classmethod
43 def from_settings(cls, settings):
44     debug = settings.getbool('SUPERFINGER_DEBUG')
45     return cls(job_dir(settings), debug)
46
47 def request_seen(self, request):
48     fp = self.request_fingerprint(request)
49     if fp in self.fingerprints:
50         return True
51     self.fingerprints.add(fp)
52     if self.file:
53         self.file.write(fp + os.linesep)
54
55 def request_fingerprint(self, request):
56     return request_fingerprint(request)
```

Recap discussion

- What are SLAs?
- What is the benefit of SLAs?
- What elements does a Data Quality SLA contain?

Building Careers
Through Education



Data Ingestion Risks

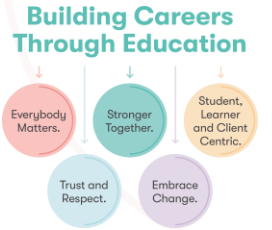
Even with properly designed SLAs, data ingestion risks will still exist.

These include:

- Ingestion system failing and needing troubleshooting (SLA will specify timelines and escalation).
- Ingestion system not behaving optimally and falling below sustainable standards.
- Security vulnerabilities of the ingestion system.
- Ingested data quality deteriorating systematically in an unplanned manner.
- PII and sensitive data being ingested and handled improperly.
- Unused ingested data may keep clogging up our infrastructure.
- Data ingestion frequency may be inadequate due to changing environment.
- Ingested data having licensing restrictions that changed or are not properly managed.
- Upstream schema changes breaking downstream systems and metadata becoming stale.

All of these risks should be included in the risk register, monitored and reported on.

We will look at them in more detail in today's webinar.



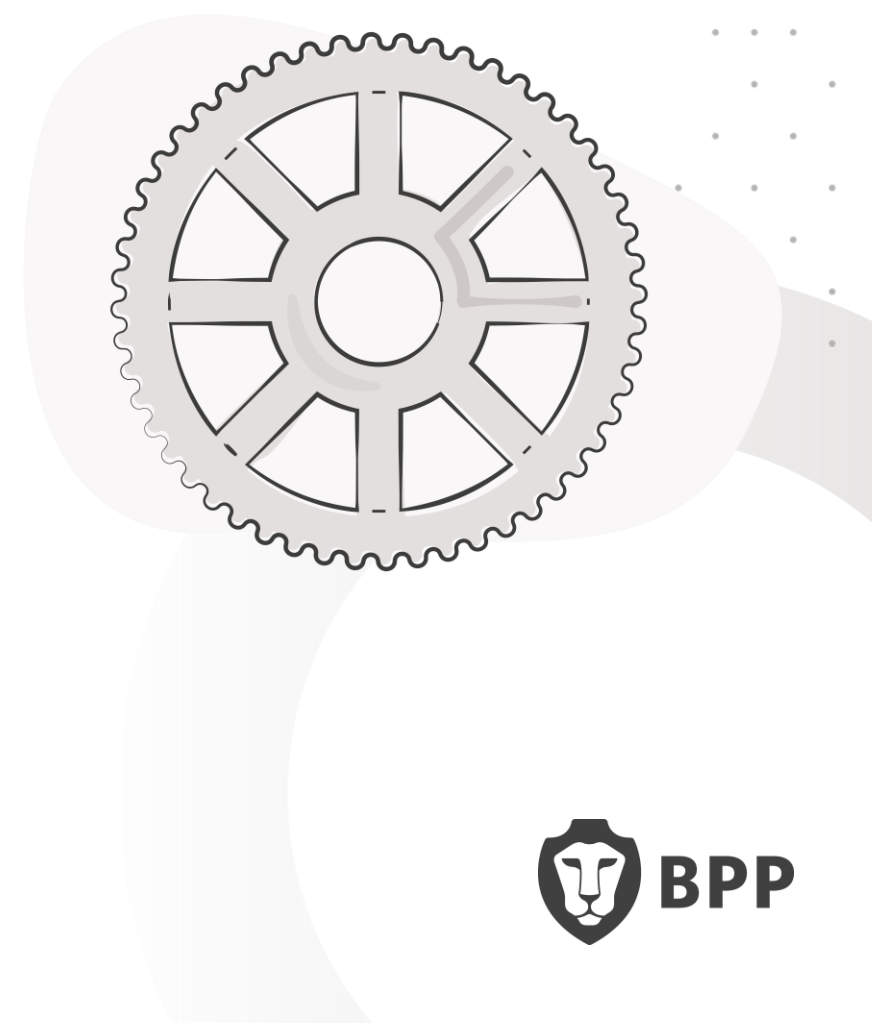
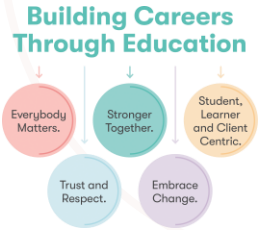
Troubleshooting practical



Your tutor will now walk you through a worked example.



From 10m30s



Lessons learned

Not understanding the intricacies of your chosen programming language is a risk – how would you mitigate it?

- Exhaustive testing, code reviews, training.

What are some practical troubleshooting steps to take in this scenario?

- Focus, and document which steps you tried that didn't work (so that you don't repeat them).
- Feel comfortable asking others for help.
- When you fix a problem, reflect on how to avoid similar problems in future.

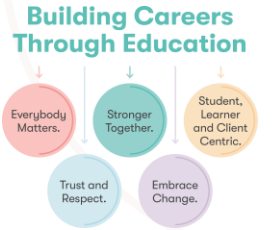
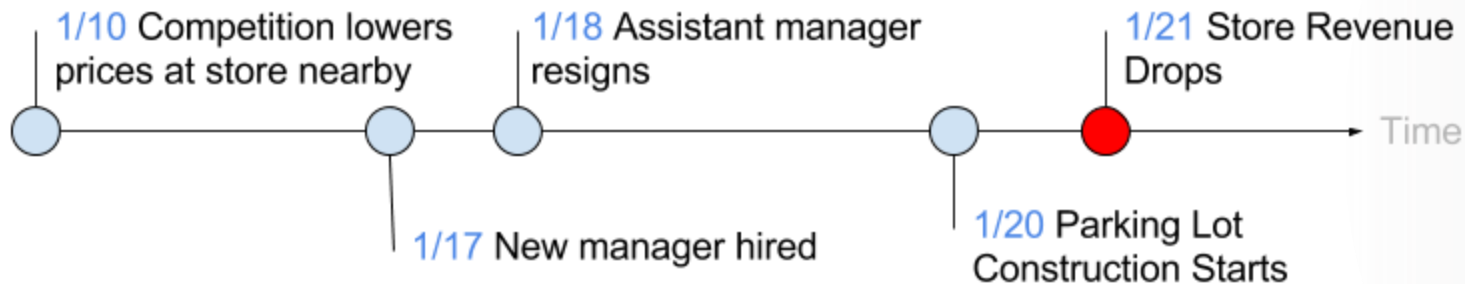
Building Careers
Through Education



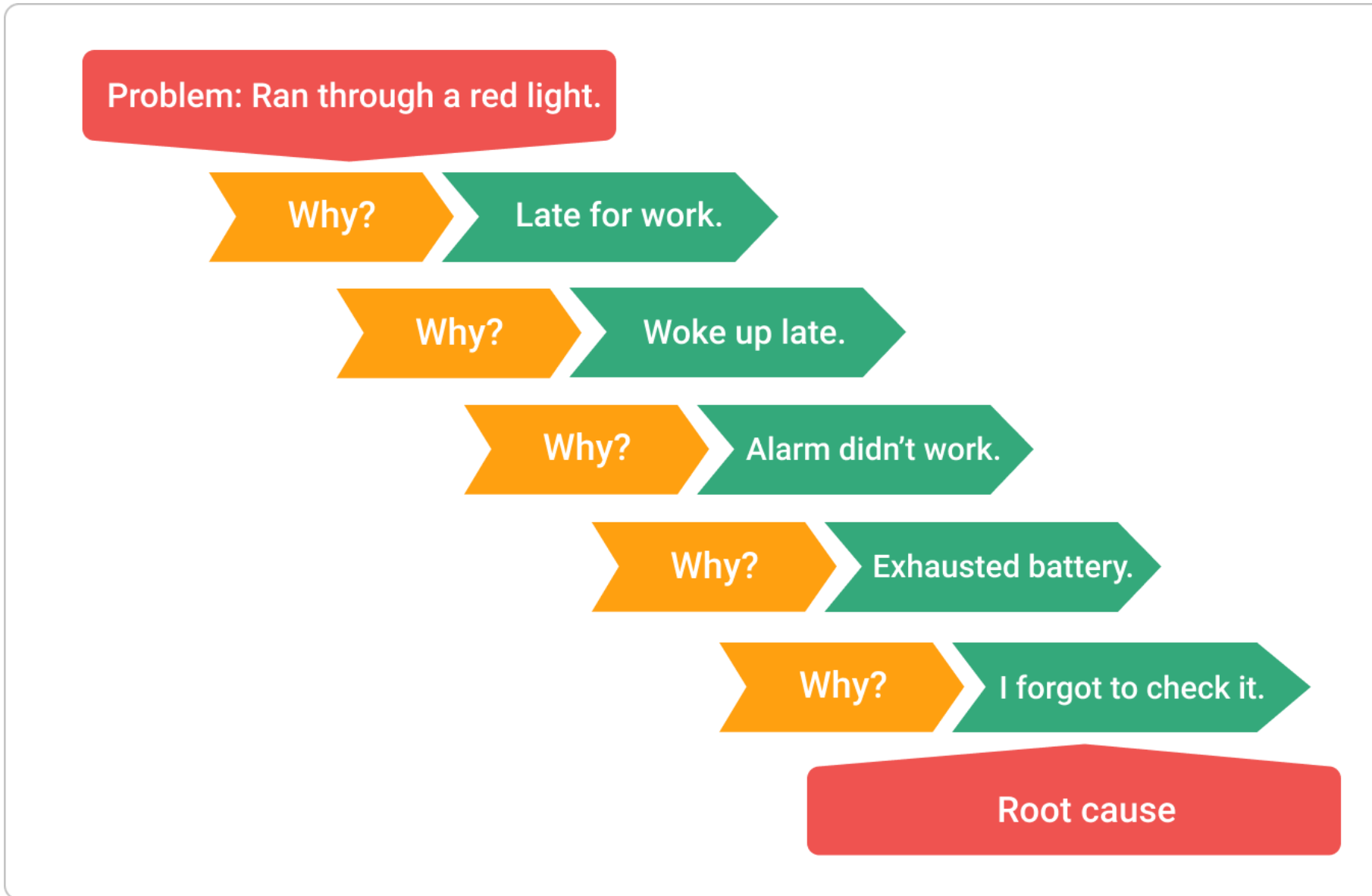
Root cause analysis

Root cause analysis normally comprises of at least the following steps:

- Identifying Contributing Factors.
- Organising Factors:
 - Sorting on a timeline.
 - Estimating the likelihood and/or severity.
- Classifying Factors:
 - Correlated Results.
 - Unrelated Factors.
 - Contributing Factors.
 - Root Cause.



The 5 Whys



Building Careers
Through Education



Net-zero and sustainability data risks

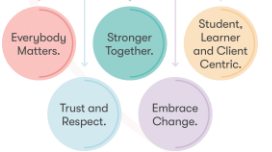
Accurate data plays a crucial role in achieving sustainability by enabling businesses to track progress, identify areas for enhancement, make informed decisions and guide resource usage.

Data engineering is a crucial enabler of sustainability and green technologies. It empowers organisations to harness the power of data to optimise operations, reduce environmental impact, and drive efficiency in eco-friendly solutions

Some of the key net-zero and sustainability data risks include:

- Poorly designed data collection and ingestion (e.g. not taking into account data from sources within green tech infrastructure).
- Poorly designed data pipelines and ETL processes.
- Lack of monitoring and/of optimisation.
- Inadequate predictive analytics.
- Lack of automation.
- Lack of proper data management such as minimisation.
- Lack of proper data optimisation, such as compression.

Building Careers
Through Education

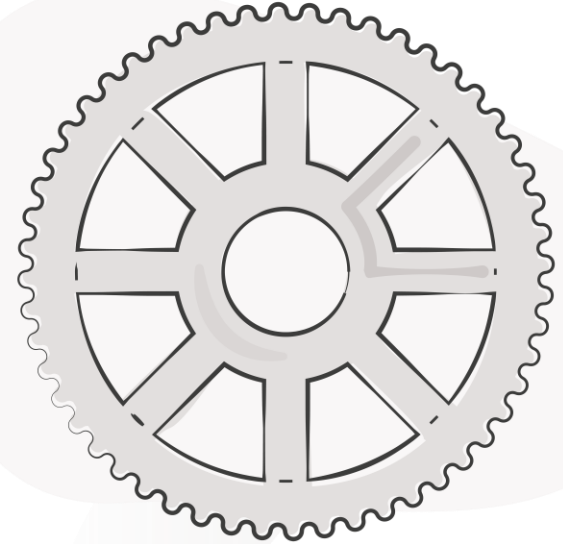


Net-zero and sustainability practical



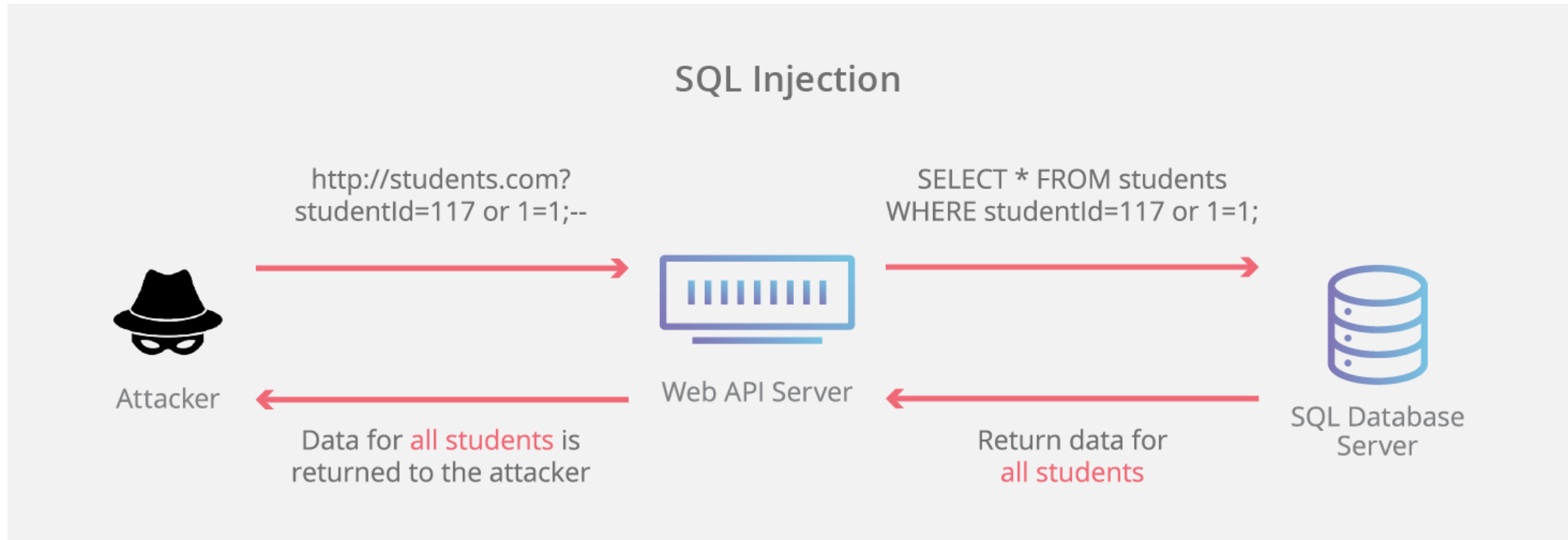
Your tutor will now walk you through a worked example.

Building Careers
Through Education



Ingestion-specific cybersecurity risks

Building Careers
Through Education



Data quality improvement

Building Careers
Through Education



Data Quality Improvement Cycle

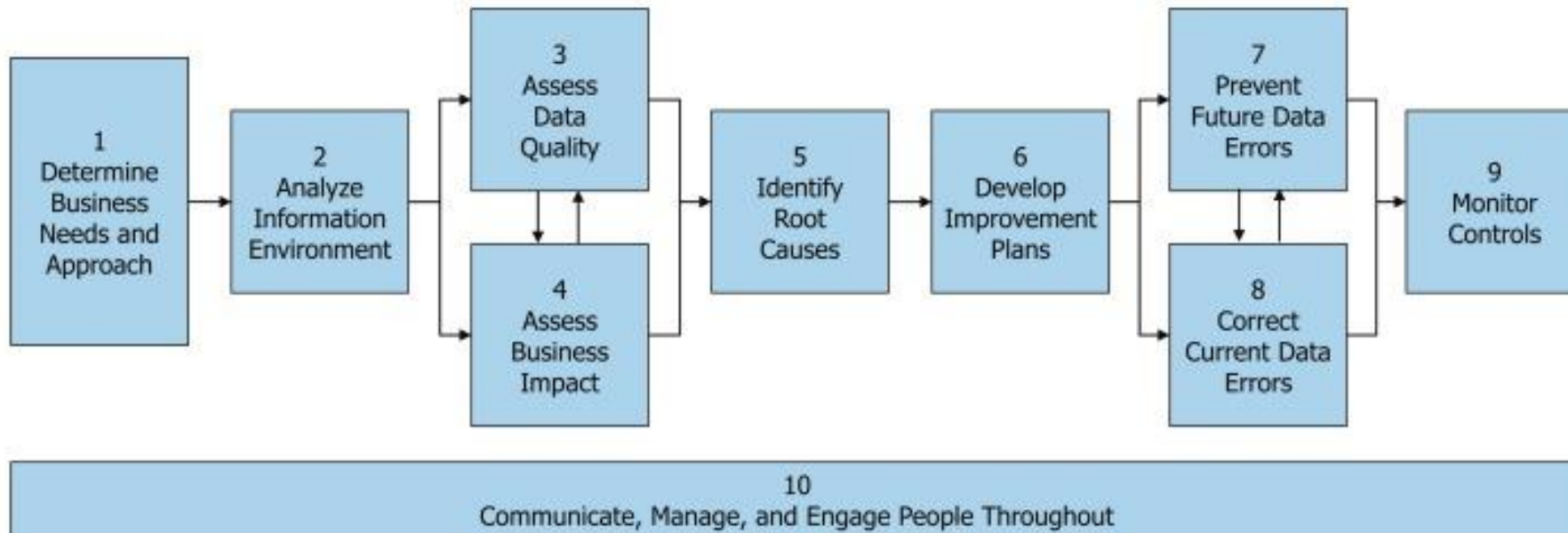
Assessment

Awareness

Action

The Ten Steps Process

Ten Steps to Quality Data and Trusted Information™



Note: Iterate among the steps as needed

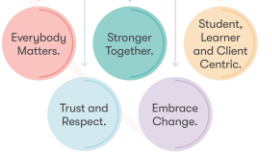
Copyright © 2005, 2020 Danette McGilvray, Granite Falls Consulting, Inc. www.gfalls.com. v10-20



Unused data and archiving



Building Careers
Through Education



Risk register

DATA RISK REGISTER TEMPLATE

RISK ID NO.	RISK DESCRIPTION	CATEGORY	RECURRENCE	IMPACT DESCRIPTION	IMPACT LEVEL	PROBABILITY LEVEL	PRIORITY LEVEL	MITIGATION OR CONTROL STRATEGY	OWNER
	Brief summary of the risk.	Select from dropdown.	ONGOING or ONE TIME	What will happen if the risk is not mitigated or eliminated.	Rate 1 (LOW) to 5 (HIGH)	Rate 1 (LOW) to 5 (HIGH)	(IMPACT X PROBABILITY) Address highest first.	What can be done to lower or eliminate the impact or probability.	Who's responsible.
		VENDOR LOCK-IN – Data trapped on a specific platform	ONGOING		4	5	20		
		DEVICE FAILURE – Hardware crashes	ONE TIME		5	5	25		
		CORRUPTION – Inaccurate data storage			2	1	2		
		REMANENCE – Residual data that remains after attempting to delete it			3	3	9		
		COMPLIANCE – Issues with laws and regulations							
		UNUSED DATA – Stored data that is never accessed							
		INTEGRATION – Data transference process failure							
		AVAILABILITY – Data not accessible by employees or customers							
		OTHER							


PROBABILITY

5	5	10	15	20	25
4	4	8	12	16	20
3	3	6	9	12	15
2	2	4	6	8	10
1	1	2	3	4	5
	1	2	3	4	5

IMPACT



Risk reporting

 MetaGovernance The Data Integrity Experts		QUICK ASSESSMENT								© MetaGovernance, Inc. 2021		
		Risk Management Activities <i>Apply performance risk rating from 1 to 5 with 5 = very mature and comprehensive</i>								Risk Weighting <i>Apply score from 1 to 5 with 1 = very low risk and 5 = very high risk</i>		
<div> <div></div> <div> Risk Management Activities (AB 2020-06) </div> </div>		Responsibilities are established	Policies, procedures, and standards are in place and followed	Management and board oversight is defined and executed	Activities follow procedures, are completed in a timely manner	Risks and issues are identified, metrics are defined and monitored	Risks and issues are reported to responsible parties, ERM, Executive, and Board	Risks and issues are mitigated on a timely basis	Average Score	Likelihood of a risk event	Potential impact of a risk event	Score
Data Management Functions (AB 2016-04)	Activities within this function											
Governance, Metadata	Data categorization (grouping), stakeholder assignments, glossary and definitions, coordination, mapping (terms,								0.00			0
User Developed Applications and Models	Identification, categorization, risk weighting, mitigating control implementation, elimination								0.00			0
Architecture, Modeling, Integration, Design	Mapping from business glossary to data sources, application data mapping, lineage, relationship mapping								0.00			0
Storage, Availability, Operations	Business continuity planning includes data, life cycle management, data base management, storage, backup and recovery; change control includes data testing, SAAS included								0.00			0
Data Security	Security classification, encryption, access control, monitoring								0.00			0
Documents and Content	Security classification and retention for unstructured data, e-discovery, crown jewel data loss prevention (DLP)								0.00			0
Master, Reference, Warehousing,	Information sources, known copies, lineage, ease of access to data, data modeling, business intelligence and reporting								0.00			0
Data Quality	Accuracy, consistency, availability, timeliness								0.00			0
SCORE		0.00	0.00	0.00	0.00	0.00	0.00	0.00				

Building Careers Through Education



Practical scenario – discussion activity

1. In your breakout rooms read the following 3 scenarios:

<https://www.precisely.com/blog/data-quality/improve-data-quality-3-examples>

2. And then read the following strategies for improving data quality (scroll down within the document):

<https://lakefs.io/data-quality/improve-data-quality/>

3. Discuss which strategy would you use for each scenario to mitigate data ingestion risks specifically.

Building Careers
Through Education



Module consolidation

Introduction to Data Collection and Ingestion	<ol style="list-style-type: none"> 1. Justify the importance of automation in data collection and ingestion 2. Evaluate common data cleaning techniques 3. Recognise the steps required to pre-process data for machine learning purposes 4. Demonstrate practical data collection skills
Heterogeneous Data Ingestion Patterns	<ol style="list-style-type: none"> 1. Explain the key design and architecture considerations of heterogeneous data ingestion systems 2. Report on the usefulness of heterogeneous ingestion patterns for business use cases 3. Demonstrate practical application of file ingestion combining multiple sources of data

APIs and Microservices for Data Engineers	<ol style="list-style-type: none"> 1. Explain the benefits of microservices architectures in data engineering 2. Demonstrate ability to use an API to ingest data 3. Design a data ingestion architecture using APIs and microservices
Data Ingestion Risks	<ol style="list-style-type: none"> 1. Explain the benefits of setting up SLAs around collecting and ingesting data 2. Explain how you would manage and report on data risks in a business environment 3. Apply strategies for collecting and ingesting PII and sensitive data 4. Evidence data quality improvement practices in practical scenarios 5. Compress ingested data in Python to reduce sustainability risks and support net-zero goals

Building Careers Through Education

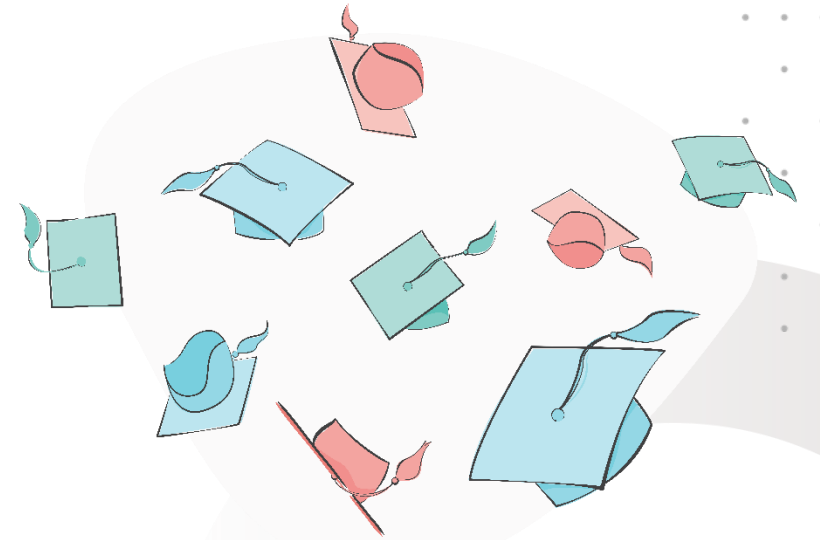


Key Learning Summary

The key takeaways from this session are as follows:

- **Benefits of SLAs in Data Collection and Ingestion:** SLAs help ensure timely data collection, maintain data quality, and address potential risks.
- **Managing Data Risks:** Effective risk management involves identifying, monitoring, and reporting on data ingestion risks, including security vulnerabilities and data quality issues.
- **Strategies for Handling PII and Sensitive Data:** Implementing robust strategies for collecting and ingesting PII and sensitive data is essential for compliance and data protection.
- **Data Quality Improvement Practices:** Regularly applying data quality improvement practices helps maintain high data standards and supports business objectives.
- **Sustainability in Data Ingestion:** Compressing ingested data and optimizing data pipelines can reduce sustainability risks and support net-zero goals.

Building Careers
Through Education





Thank you

**Do you have any questions,
comments, or feedback?**

