Submit Turnitin Assignment

Learn Help

Submit: Single File Upload

Congratulations - your submission is complete! This is your digital recei

copy of this receipt from within the Document Viewer.

« Page 1 »

Author:
Daniel Bee

Assignment title:
PItDS Assessment REPORT

Submission title:
B173513

File name:
B173513_pdf.pdf

File size:
425.49K

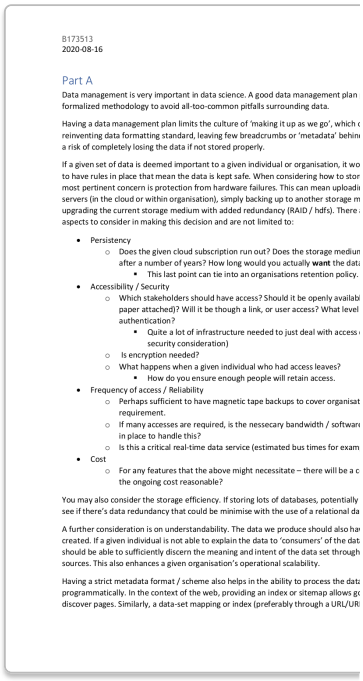
Page count:
8

Word count:
1809

Character count:
9057

Submission date:
2020年08月17日 04:11AM (UTC+0100)

Submission ID:
132467915



B173513
2020-08-16

Part A

Data management is very important in data science. A good data management plan
formalized methodology to avoid all-too-common pitfalls surrounding data.

Having a data management plan limits the culture of 'making it up as we go', which
reinventing data formatting standard, leaving few breadcrumbs or 'metadata' behin
a risk of completely losing the data if not stored properly.

If a given set of data is deemed important to a given individual or organisation, it wo
to have rules in place that mean the data is kept safe. When considering how to stor
most pertinent concern is protection from hardware failures. This can mean uploadi
servers (in the cloud or within organisation), simply backing up to another storage m
upgrading the current storage medium with added redundancy (RAID / hdfs). There
aspects to consider in making this decision and are not limited to:

- Persistency
  - Does the given cloud subscription run out? Does the storage mediu
    after a number of years? How long would you actually **want** the dat
    - This last point can tie into an organisations retention policy.
- Accessibility / Security
  - Which stakeholders should have access? Should it be openly availabl
    paper attached)? Will it be though a link, or user access? What level
    authentication?
    - Quite a lot of infrastructure needed to just deal with access
      security consideration)
  - Is encryption needed?
  - What happens when a given individual who had access leaves?
    - How do you ensure enough people will retain access.
- Frequency of access / Reliability
  - Perhaps sufficient to have magnetic tape backups to cover organisat
    requirement.
  - If many accesses are required, is the nessecary bandwidth / softwar
    in place to handle this?
  - Is this a critical real-time data service (estimated bus times for exam
- Cost
  - For any features that the above might necessitate – there will be a c
    the ongoing cost reasonable?

You may also consider the storage efficiency. If storing lots of databases, potentially
see if there's data redundancy that could be minimise with the use of a relational da

A further consideration is on understandability. The data we produce should also ha
created. If a given individual is not able to explain the data to 'consumers' of the dat
should be able to sufficiently discern the meaning and intent of the data set through
sources. This also enhances a given organisation's operational scalability.

Having a strict metadata format / scheme also helps in the ability to process the data
programmatically. In the context of the web, providing an index or sitemap allows g
discover pages. Similarly, a data-set mapping or index (preferably through a URL/UR

Return to assignment list