WQD7007 Big Data Management

Data Provenance
and
Data Trustworthiness

Data Provenance

- A formal representation of computational processes.
  - Documented precise descriptions of the **origin** of data, the **transformations** that have been applied to those data, and the **implications of the results**.
  - It **should be collected automatically** in a manner amenable to automated reasoning, so that data origin, data processing, and results presentation are communicated to a user in an intelligible manner.

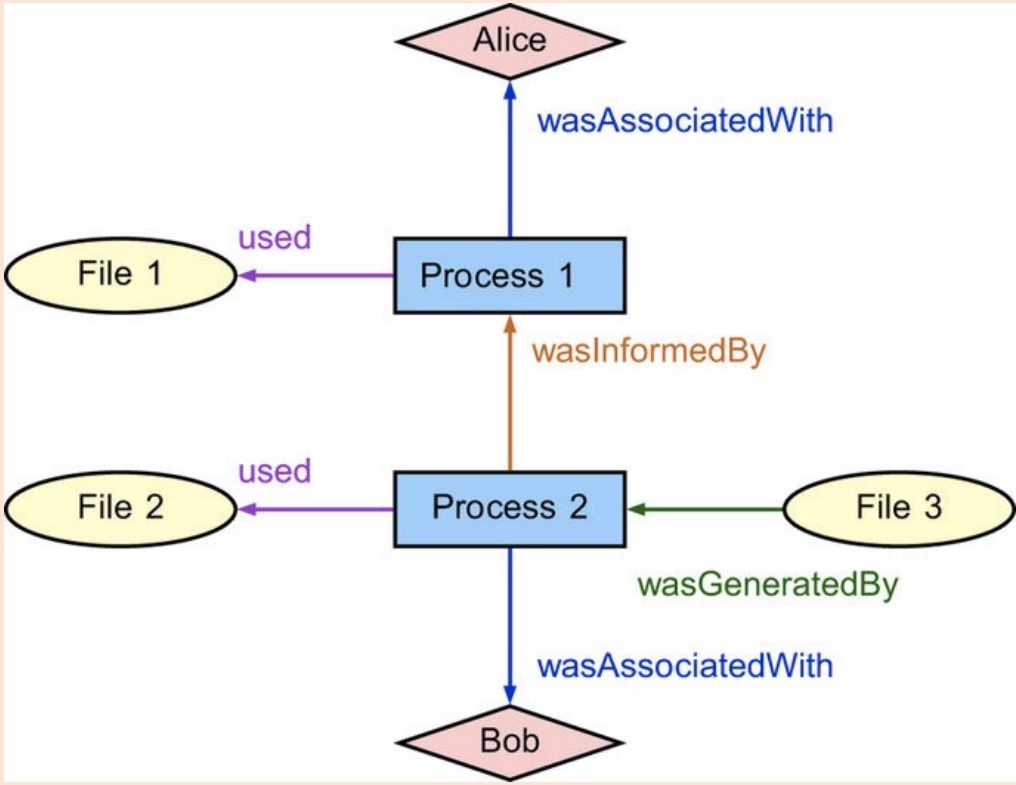
Source: https://www.nature.com/articles/sdata2017114

Data Provenance

- Provenance data is most frequently represented as a directed, acyclic graph.
    - Interactions are recorded as a set of *edges* that relate data-items, transformations (computations), and persons or organizations associated with the data, all represented as *vertices*

Data Provenance

- In this example, two processes (Process 1) and (Process 2), use the data from the inputs File 1 and File 2, respectively.
- The processes are associated respectively with the users Alice and Bob, respectively.
  - Process 1 informed (transferred information to) Process 2, which generated the output File 3.



Data Provenance

- Exploiting data provenance is a multi-stage process.

1. First, it involves the ***capture*** of data provenance during code execution.
2. Next, the provenance must be ***stored*** in an efficient manner.
3. Last, the provenance is ***queried and analysed*** either by machines (algorithmically) or by humans, most frequently through visualization.

Data Provenance - capture

- Provenance capture can be divided into two broad categories: observed and disclosed.
  - **Disclosed provenance** consists of modifying an existing application so that it publishes the provenance resulting from its execution.
    - One example of disclosed provenance is the Earth System Science Workbench, used to process satellite imagery.
  - **Observed provenance** consists of modifying the system on top of which the computation runs, so that it systematically and automatically records how data are generated.
    - PASS, which captures provenance in the operating system, is an observed provenance system. It produces a record of the execution of unmodified programs that run on top of it.

Data Provenance - Storage

- Numerous systems have been developed over the years to accomplish this.
  - Some are domain-specific, whereas others have been intended for more general application.
  - For example, the **Core Provenance Library** provides an interface between provenance generating applications and various database back-ends. It enables the integration of provenance information from diverse sources into a coherent whole.
  - Other issues, such as the scale of provenance generated by large scale systems, are being addressed using Big Data storage.

Data Provenance - Analysis

- The last aspect concerns the analysis, query and use of the provenance data.
  - For example, visualization tools that present provenance data in an intelligible manner have been created by projects, such as **Orbiter or VisTrails**.
- Another use of provenance is to render the analytical process transparent. By examining provenance records, one can learn **how a team went from the raw collected data to the published results**.
  - Tools such as ReproZip have been built to automatically reproduce computational environments.

Data Provenance - Analysis

- Others have envisioned using these data to produce executable papers to allow readers and reviewers to **repeat a computational experiment** or conduct related experiments.
  - Additionally, they can be used to verify a claim or test new hypotheses with less engineering effort.
  - Examples of executable papers from the Association for Computing Machinery Special Interest Group on Management of Data 2008 to 2011 conferences are available

Data Provenance – Importance in Big Data Debugging

- Massive Scale
- Unstructured data
- Long Runtime
- Complex Platform

- These challenges are overcome by careful data provenance design.

Data Trustworthiness

- In order for analysts and decision makers to produce accurate analysis and make effective decisions and take actions, data must be **trustworthy**
  - Important in data sharing, situation assessment, multi-sensor data integration and numerous other functions to support decision makers and analysts
  - E.g. critical decision making in traffic monitoring systems and power grid management systems, where a large amount of data that can convey important information is collected from distributed sources

Source: https://pdfs.semanticscholar.org/1425/3cd72cccf09d3bee3ffc704655dcf253b482.pdf

Ways to ensure data trustworthiness

1. A mechanism for associating **confidence values** with data in the database.
   - A confidence value is a numeric value ranging from 0 to 1 and indicates the trustworthiness of the data.
   - Confidence values can be generated based on various factors, such as the **trustworthiness of data providers** and **the way in which the data has been collected.**
2. The notion of **confidence policy**, indicating which confidence level is required for certain data when used in certain tasks.

Ways to ensure data trustworthiness

3. The **computation of the confidence values of a query result**s based on the confidence values of each data item and lineage propagation techniques.
4. The fourth element is **a set of strategies for incrementing the confidence of query results at query processing time**.
   - Such element is a crucial component in that it makes possible to return query results meeting the confidence levels stated by the confidence policies.

Data Trustworthiness – Assigning Confidence Levels to Data

- Based on the trustworthiness of data provenance
  - From **which sources** did the data originate from?
  - How **trustworthy** are such data sources?
  - Which **entities** (applications or systems) handled the data? Are these entities trustworthy?

Data Trustworthiness – Assigning Confidence Levels to Data

- A Data Provenance Trust Model measure three different aspects that may affect data trustworthiness: data similarity, data conflict, and path similarity

  - **Data similarity** in this model refers to the likeness of different items. Similar items are considered as supportive to each other.
  - **Data conflict** refers to inconsistent descriptions or information about the same entity or event.
  - **Path similarity** models how similar are the paths followed by two data items from the sources to the destination.

Data Trustworthiness – Policies

- Regulating the use of the data according to requirements concerning data confidence levels.
  - motivated by the observation that the required level of data trustworthiness **depends on the purpose** for which the data have to be used.
  - For example, for tasks which are **not critical** to an organization, like computing a statistical summary, data with a medium confidence level may be sufficient, whereas when an individual in an organization has to make a **critical** decision, data with high confidence are required.

Data Trustworthiness – Policies

- A policy in the confidence policy model specifies the **minimum confidence** that has to be assured for certain data, depending on the user accessing the data and the purpose the data access.
  - In its essence, a confidence policy contains three components:
    1. **a subject specification**, denoting a subject or set of subjects to whom the policy applies;
    2. **a purpose specification**, denoting why certain data are accessed;
    3. **a confidence level**, denoting the minimum level of confidence that has to be assured by the data covered by the policy when the subject to whom the policy applies requires access to the data for the purpose specified in the policy

Data Trustworthiness – Policy Complying Query Evaluation

- A critical issue in enforcing confidence policies in the context of query processing is that **some of the query results may be filtered out due to confidence policy violation**.
  - Therefore a user **may not receive enough data to make a decision** and he may want to improve the data quality.
  - One way is to dynamically incrementing the data confidence level. Such approach selects an optimal strategy which determines **which data should be selected and how much the confidence should be increased** to comply with the confidence level stated by the policies.

Data Trustworthiness – Policy Complying Query Evaluation

- A Policy Complying Query Evaluation System consists of four main components:
  1. **query evaluation,**
  2. **policy evaluation,**
  3. **strategy finding, and**
  4. **data quality improvement.**

Data Trustworthiness – Policy Complying Query Evaluation

- Data flow within the system is first elaborated. Initially, each base tuple is assigned a confidence value by the confidence assignment component

  - A user inputs query information in the form , where *Q* is a normal SQL query, *purpose* indicates the purpose for which the data returned by the query will be used, and *perc* indicates percentage of results that the user expects to receive after the confidence policy enforcement.

Data Trustworthiness – Policy Complying Query Evaluation

- The **query evaluation** component then computes the results of Q and the confidence level of each tuple in the result based on the confidence values of base tuples.
- The intermediate results are sent to the **policy evaluation** component.
- The policy evaluation component selects the confidence policy associated with the role of the user who issued Q and checks each tuple in the query result according to the selected confidence policy.
- Only the results with confidence value higher than the threshold specified in the confidence policy are immediately returned to the user.

Data Trustworthiness – Policy Complying Query Evaluation

- If less than *perc* of the results satisfy the confidence policy, the strategy finding component is invoked to devise an optimal strategy for increasing the confidence values of the base tuples and report the cost of such strategy to the user.

- If the user agrees about the cost, the **strategy finding** component will inform the **data quality improvement** component to take actions to improve the data quality and then update the database.

- Finally, new results will be returned to the user.