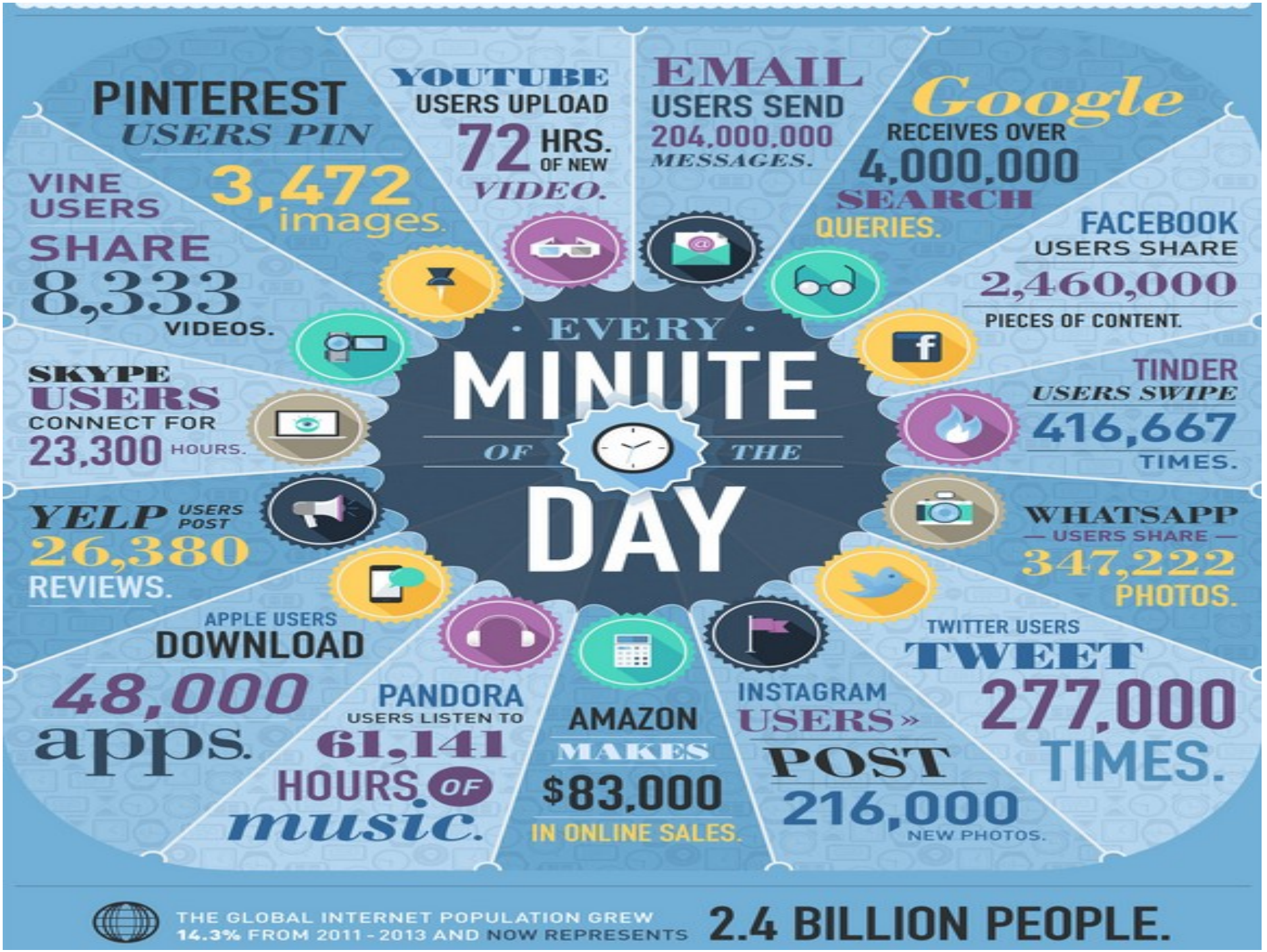
Data Lineage for distributed data systems

Sravani Kamisetty



http://newstex.com/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/

Challenges in Big Data Debugging

- Massive scale
- Unstructured data
- Long runtime
- Complex platform

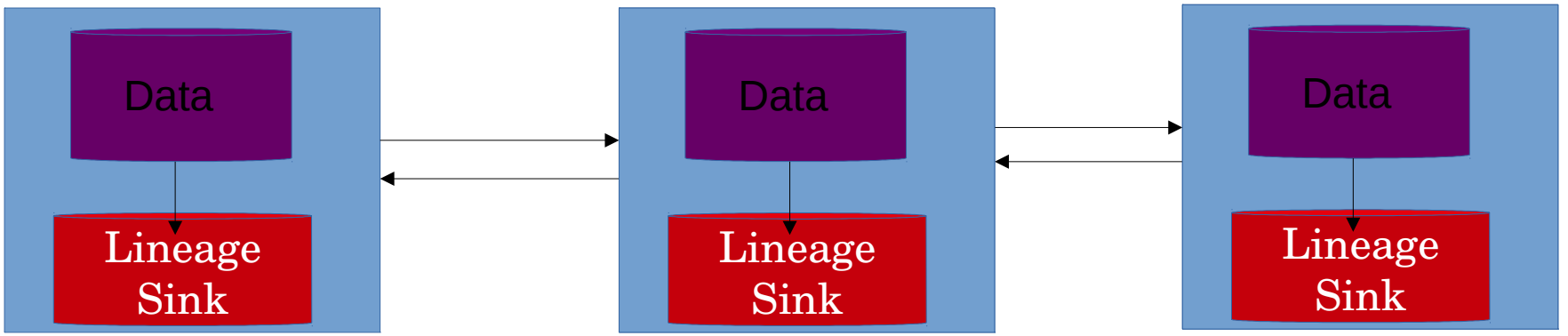
What does Data Lineage mean?

- Data life cycle that includes the data's origins and where it moves over time.
- Describes what happens to data as it goes through diverse processes.

Lineage Capture

- Lineage is captured as a triplet of form {I, T, o}, where I is the set of inputs to T used to derive o.

- "Which outputs were produced by an input i on operator T?" and "Which inputs produced output o on operator T?"



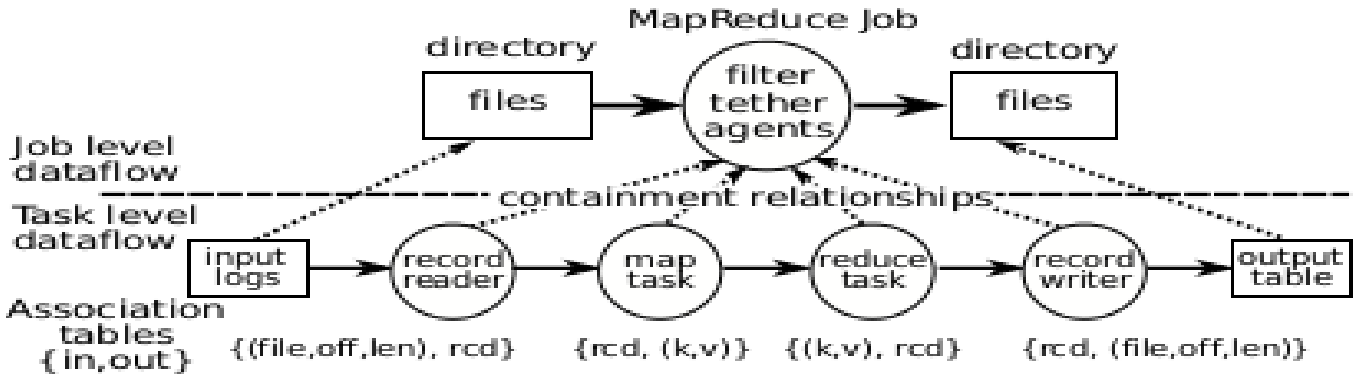
Actors & Associations

- An actor is an entity that transforms data.

  Eg: individual map and reduce operators, a MapReduce job or an entire dataflow pipeline.

- Association is the relationship between the input and output i.e the operation

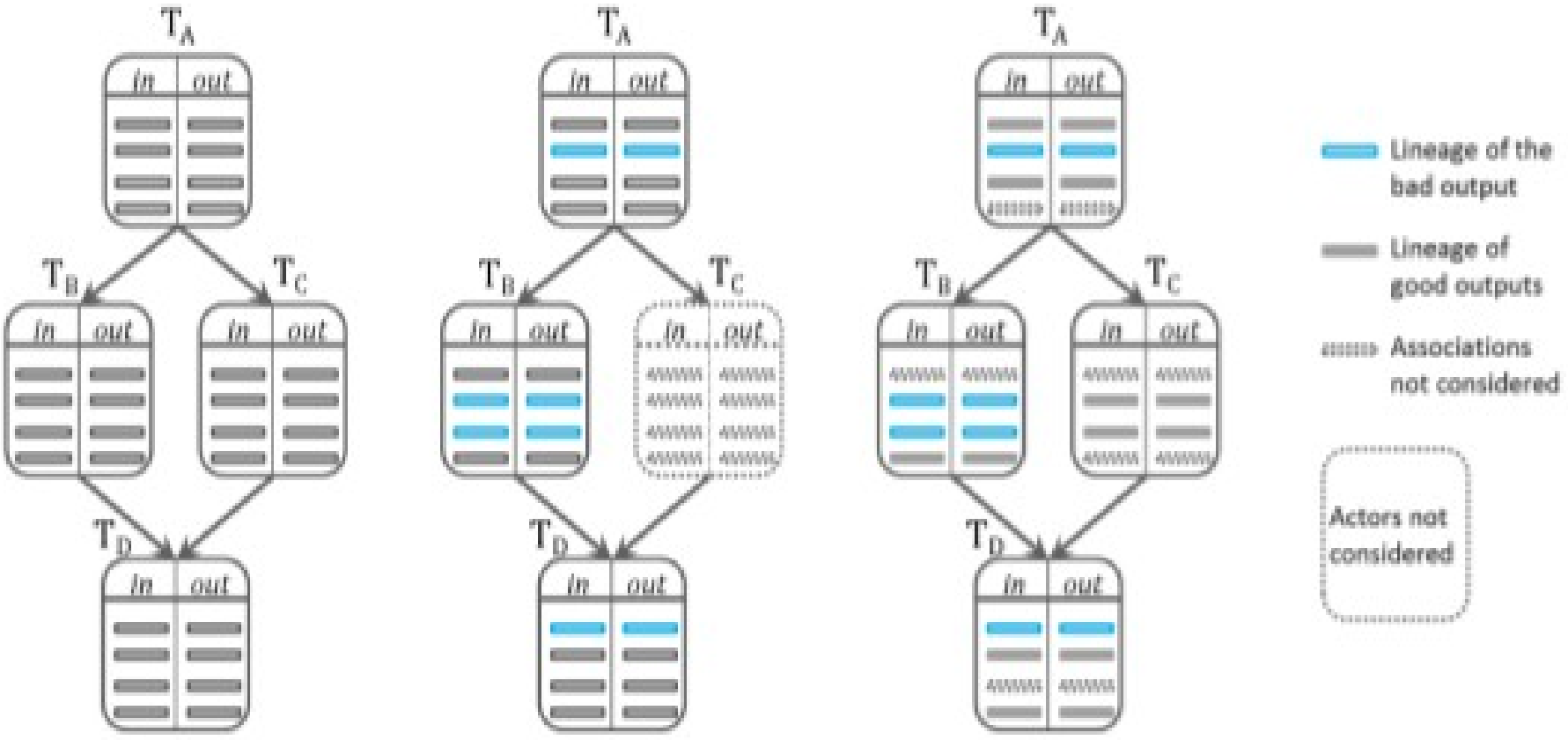
- Associations are stored in an association table on each peer.



Data-flow reconstruction

- A graph of the actors is built to represent the data-flow
- Explicitly specified links – Actor is aware of the upstream/downstream actor
- Logically inferred links – Infer links from archetype of the dataflow
- Implicit links through dataset sharing

Tracing & Exclusive Replay



Challenges

- Load balancing
- Fault Tolerance
- Scalability

References

[1] Dionysios Logothetis, Soumyarupa De, and Kenneth Yocum. 2013. Scalable lineage capture for debugging DISC analytics. In Proceedings of the 4th annual Symposium on Cloud Computing (SOCC '13). ACM, New York, NY, USA, , Article 17 , 15 pages.

[2] De, Soumyarupa. (2012). Newt : an architecture for lineage -based replay and debugging in DISC systems. UC San Diego: b7355202. Retrieved from: https://escholarship.org/uc/item/3170p7zn

[3] Hao Fan and Ra Poulovassilis. Using schema transformation pathways for data lineage tracing. In In BNCOD, pages 133–144. Springer, 2005.

Thank you!