# Data Skew

Luc Bouganim

# Data Skew

Luc Bouganim
INRIA Saclay Île de France & UVSQ Versailles, France

**Luc Bouganim**
**Email:** luc.bouganim@inria.fr

## Synonyms

Biased distribution; Non-uniform distribution

## Definition

Data skew primarily refers to a non uniform distribution in a dataset. Skewed distribution can follow common distributions (e.g., Zipfian, Gaussian, Poisson), but many studies consider Zipfian [ *1*] distribution to model skewed datasets. Using a real bibliographic database, [ *2*] provides real-world parameters for the Zipf distribution model. The direct impact of data skew on parallel execution of complex database queries is a poor load balancing leading to high response time.

## Key Points

Walton et al. [ *3*] classify the effects of skewed data distribution on a parallel execution, distinguishing *intrinsic skew* from *partition skew*. Intrinsic skew is skew inherent in the dataset (e.g., there are more citizens in Paris than in Waterloo) and is thus called *Attribute value skew (AVS)*. Partition skew occurs on parallel implementations when the workload is not evenly distributed between nodes, even when input data is uniformly distributed. Partition skew can further be classified in four types of skew. *Tuple placement skew (TPS)* is the skew introduced when the data is initially partitioned (e.g., with range partitioning). *Selectivity skew (SS)* is introduced when there is variation in the selectivity of select predicates on each node. *Redistribution skew (RS)* occurs in the redistribution step between two operators. It is similar to TPS. Finally *join product skew (JPS)* occurs because the join selectivity may vary between nodes.

## Cross-References

Query Load Balancing in Parallel Database Systems

# Recommended Reading


1. Zipf GK. Human behavior and the principle of least effort: an introduction to human ecology. Reading: Addison-Wesley; 1949.
2.. Lynch C. Selectivity estimation and query optimization in large databases with highly skewed distributions of column values. Proceedings of the 14th International Conference on Very Large Data Bases; 1988. p. 240–51.

3. Walton CB, Dale AG, Jenevin RM. A taxonomy andperformance model of data skew effects in parallel joins. Proceedings of the 17th International Conference on Very Large Data Bases; 1991. p. 537–48.