Data Lake Zones

Last updated: April 8, 2018

Raw Data Zone

- ✓ Exact copy of source data in native format (aka master dataset in the batch layer)
- ✓ Immutable to change
- ✓ History retained indefinitely
- ✓ Data access is highly limited to few people
- ✓ Everything downstream can be regenerated from raw

Transient/Temp Zone

- ✓ Selectively utilized
- ✓ Separation of "new data" from "raw data" to ensure data consistency
- ✓ Transient low-latency data (aka speed layer)
- ✓ Data quality validations

Master Data Zone

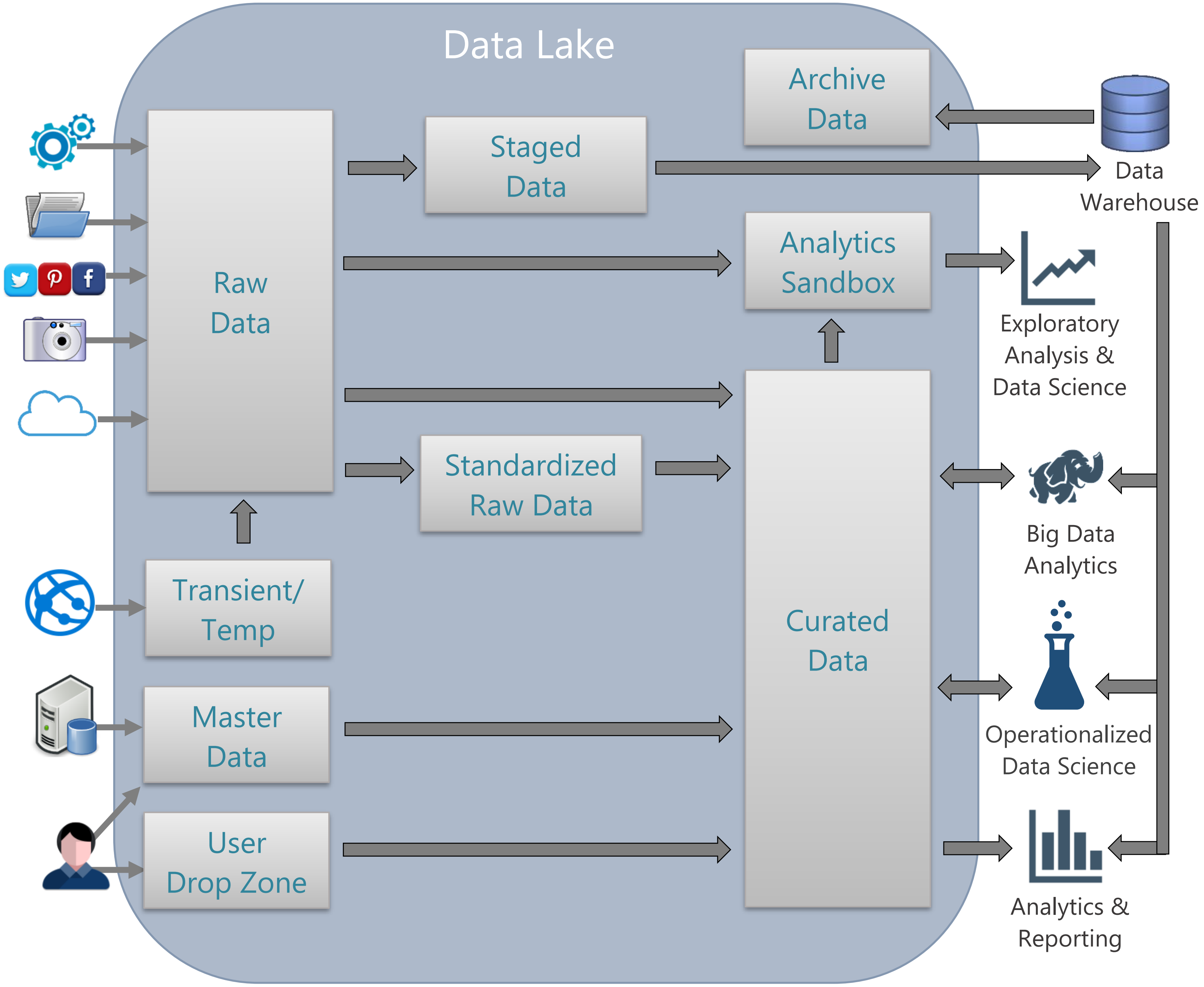
- ✓ Reference data

User Drop Zone

- ✓ Manually generated data

Staged Data Zone

- ✓ Data staged for a specific purpose or application



Metadata | Security | Governance | Information Management

Standardized Raw Data

- ✓ Raw data which varies in format or schema, such as JSON which is standardized into columns & rows (aka "semantic normalization")
- ✓ File consolidations of data (i.e., to overcome performance issues with many small files)

Archive Data Zone

- ✓ Active archive of aged data, available for querying when needed

Analytics Sandbox

- ✓ Workspace for exploratory data science & analytics
- ✓ Valuable efforts are productionized to the curated data zone

Curated Data Zone

- ✓ Cleansed and transformed data, organized for optimal data delivery (aka serving layer)
- ✓ Supports self-service
- ✓ Standard security, change management, and governance

Latest version: http://www.sqlchick.com > Presentations & Downloads page