Curating Data continued

Indexing and Cataloging

Indexing and cataloging functions occur automatically in PHEMI Central's Indexing and Cataloging Engine, making it easier and faster to find and consume data. User-defined DPFs enable deeper and more sophisticated indexing and cataloguing, and second-order indexes and graph relationships allow data analysts to quickly find and build datasets across petabytes of heterogeneous digital assets. Linking datasets with common keys makes it faster and easier to build meaningful datasets across many sources. These powerful indexing features mean that data can be accessed in milliseconds, without having to wait for MapReduce or YARN jobs to complete.

Powerful indexing features mean that data can be accessed in milliseconds without having to wait for MapReduce or YARN jobs to complete.

Schemaless Storage

PHEMI Central's data store is schemaless: both raw and curated data items are stored in a binary format that is unaffected by the source and destination schema. This approach means that organizations can quickly aggregate new data sources without costly redefinition of old schemas. Schemaless storage also permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas. Instead, PHEMI Central uses a flexible, powerful, distributed key-value store and sophisticated metadata tagging to manage, describe, and govern the data it stores. Curated digital assets derived from the raw data are linked to the original raw data, but PHEMI Central's SQL and REST interfaces abstract away from internal linkages and structures, so users and applications can focus on data use rather than data janitorial work.

Schemaless storage permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas

alt layout 1 for "Curating Data" p.2 (no overlap on boxes)

Curating Data continued

Indexing and Cataloging

Indexing and cataloging functions occur automatically in PHEMI Central's Indexing and Cataloging Engine, making it easier and faster to find and consume data. User-defined DPFs enable deeper and more sophisticated indexing and cataloguing, and second-order indexes and graph relationships allow data analysts to quickly find and build datasets across petabytes of heterogeneous digital assets. Linking datasets with common keys makes it faster and easier to build meaningful datasets across many sources. These powerful indexing features mean that data can be accessed in milliseconds, without having to wait for MapReduce or YARN jobs to complete.

Powerful indexing features mean that data can be accessed in milliseconds without having to wait for MapReduce or YARN jobs to complete.

Schemaless Storage

PHEMI Central's data store is schemaless: both raw and curated data items are stored in a binary format that is unaffected by the source and destination schema. This approach means that organizations can quickly aggregate new data sources without costly redefinition of old schemas. Schemaless storage also permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas. Instead, PHEMI Central uses a flexible, powerful, distributed key-value store and sophisticated metadata tagging to manage, describe, and govern the data it stores. Curated digital assets derived from the raw data are linked to the original raw data, but PHEMI Central's SQL and REST interfaces abstract away from internal linkages and structures, so users and applications can focus on data use rather than data janitorial work.

Schemaless storage permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas

Curating Data continued

Powerful indexing features mean that data can be accessed in milliseconds without having to wait for MapReduce or YARN jobs to complete.

Indexing and Cataloging

Indexing and cataloging functions occur automatically in PHEMI Central's Indexing and Cataloging Engine, making it easier and faster to find and consume data. User-defined DPFs enable deeper and more sophisticated indexing and cataloguing, and second-order indexes and graph relationships allow data analysts to quickly find and build datasets across petabytes of heterogeneous digital assets. Linking datasets with common keys makes it faster and easier to build meaningful datasets across many sources. These powerful indexing features mean that data can be accessed in milliseconds, without having to wait for MapReduce or YARN jobs to complete.

Schemaless Storage

PHEMI Central's data store is schemaless: both raw and curated data items are stored in a binary format that is unaffected by the source and destination schema. This approach means that organizations can quickly aggregate new data sources without costly redefinition of old schemas. Schemaless storage also permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas. Instead, PHEMI Central uses a flexible, powerful, distributed key-value store and sophisticated metadata tagging to manage, describe, and govern the data it stores. Curated digital assets derived from the raw data are linked to the original raw data, but PHEMI Central's SQL and REST interfaces abstract away from internal linkages and structures, so users and applications can focus on data use rather than data janitorial work.

Schemaless storage permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas.

alt layout 2 for "Curating Data" p.2 (boxes kept to columns, diff colours)