|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Simulation driven Materials Genomics | |
| **Vertical (area)** | | Scientific Research: Materials Science | |
| **Author/Company/Email** | | David [Skinner/LBNL/deskinner@lbl.gov](mailto:Skinner/LBNL/deskinner@lbl.gov) | |
| **Actors/Stakeholders and their roles and responsibilities** | | Capability providers: National labs and energy hubs provide advanced materials genomics capabilities using computing and data as instruments of discovery.  User Community: DOE, industry and academic researchers as a user community seeking capabilities for rapid innovation in materials. | |
| **Goals** | | Speed the discovery of advanced materials through informatically driven simulation surveys. | |
| **Use Case Description** | | Innovation of battery technologies through massive simulations spanning wide spaces of possible design. Systematic computational studies of innovation possibilities in photovoltaics. Rational design of materials based on search and simulation. | |
| **Current**  **Solutions** | **Compute(System)** | | Hopper.nersc.gov (150K cores) , omics-like data analytics hardware resources. |
| **Storage** | | GPFS, MongoDB |
| **Networking** | | 10Gb |
| **Software** | | PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, varied community codes |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Gateway-like. Data streams from simulation surveys driven on centralized peta/exascale systems. Widely distributed web of dataflows from central gateway to users. |
| **Volume (size)** | | 100TB (current), 500TB within 5 years. Scalable key-value and object store databases needed. |
| **Velocity**  **(e.g. real time)** | | High-throughput computing (HTC), fine-grained tasking and queuing. Rapid start/stop for ensembles of tasks. Real-time data analysis for web-like responsiveness. |
| **Variety**  **(multiple datasets, mashup)** | | Mashup of simulation outputs across codes and levels of theory. Formatting, registration and integration of datasets. Mashups of data across simulation scales. |
| **Variability (rate of change)** | | The targets for materials design will become more search and crowd-driven. The computational backend must flexibly adapt to new targets. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Validation and UQ of simulation with experimental data of varied quality. Error checking and bounds estimation from simulation inter-comparison. |
| **Visualization** | | Materials browsers as data from search grows. Visual design of materials. |
| **Data Quality (syntax)** | | UQ in results based on multiple datasets.  Propagation of error in knowledge systems. |
| **Data Types** | | Key value pairs, JSON, materials fileformats |
| **Data Analytics** | | MapReduce and search that join simulation and experimental data. |
| **Big Data Specific Challenges (Gaps)** | | HTC at scale for simulation science. Flexible data methods at scale for messy data. Machine learning and knowledge systems that integrate data from publications, experiments, and simulations to advance goal-driven thinking in materials design. | |
| **Big Data Specific Challenges in Mobility** | | Potential exists for widespread delivery of actionable knowledge in materials science. Many materials genomics “apps” are amenable to a mobile platform. | |
| **Security & Privacy**  **Requirements** | | Ability to “sandbox” or create independent working areas between data stakeholders. Policy-driven federation of datasets. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | An OSTP blueprint toward broader materials genomics goals was made available in May 2013. | |
| **More Information (URLs)** | | http://www.materialsproject.org | |
| **Note:** <additional comments> | | | |