This extends Big Data Ogres to Big Data and Simulation (HPC) Convergence Diamonds

There are still 4 views but facets are changed somewhat

Facet n (without D or M) refers to a facet of system including data and/or model

Facet nD is a Data only facet

Facet nM is a Model only facet

The Simulation Processing Diamonds are based on Berkeley Dwarfs and NAS Parallel Benchmarks

Remember Big Data has a model, so there are model diamonds for big data – they describe analytics. For some facets there are separate data and model facets. A good example in “Diamond Micropatterns or Execution Features” is that 4D is Data Volume and 4M Model size

|  |  |  |
| --- | --- | --- |
| Big Data and Simulation (HPC) Convergence Diamonds | | |
| Facet and View | | Comments |
| Problem Architecture View of Diamonds (Meta or MacroPatterns); Nearly all Data and Model | | |
| 1 | Pleasingly Parallel | As in BLAST, Protein docking. Includes Local Analytics or Machine Learning – ML or filtering pleasingly parallel, as in bio-imagery, radar images (pleasingly parallel but sophisticated local analytics) |
| 2 | Classic MapReduce | Search, Index and Query and Classification algorithms like collaborative filtering. |
| 3 | Map-Collective | Iterative maps + communication dominated by “collective” operations as in reduction, broadcast, gather, scatter. Common datamining pattern but also seen in simulations |
| 4 | Map Point-to-Point | Iterative maps + communication dominated by many small point to point messages as in graph algorithms and simulations |
| 5 | Map Streaming | Describes streaming, steering and assimilation problems |
| 6 | Shared memory (as opposed to distributed parallel algorithm) | Corresponds to problem where shared memory implementations important. Tend to be dynamic and asynchronous |
| 7 | SPMD | Single Program Multiple Data, common parallel programming feature |
| 8 | Bulk Synchronous Processing BSP | well-defined compute-communication phases |
| 9 | Fusion | Full applications often involves fusion of multiple methods. Only present for composite Diamonds |
| 10 | Dataflow | Important application features often occurring in composite Diamonds |
| 11M | Agents | Used in areas like epidemiology (swarm approaches) |
| 12 | Orchestration (workflow) | All applications often involve orchestration (workflow) of multiple components |
|  | | |
| Diamond Micropatterns or Execution Features | | |
| 1 | Performance Metrics | Result of Benchmark |
| 2 | Flops per Byte (Memory or I/O). Flops per watt (power). | I/O Not needed for “pure in memory” benchmark. |
| 3 | Execution Environment | Core libraries needed: matrix-matrix/vector algebra, conjugate gradient, reduction, broadcast; Cloud, HPC, threads, message passing etc. Could include details of machine used for benchmarking here |
| 4D | Data Volume | Property of a Diamond Instance. Benchmark measure |
| 4M | Model Size |
| 5D | Data Velocity | Associated with streaming facet but value depends on particular problem. Not applicable to model |
| 6D | Data Variety | Most useful for composite Diamonds. Applies separately for model and data |
| 6M | Model Variety |
| 7 | Veracity | Most problems would not discuss but potentially important |
| 8M | Communication Structure | Interconnect requirements; Is communication BSP, Asynchronous, Pub-Sub, Collective, Point to Point? Distribution and Synch |
| 9D | D=Dynamic or S=Static Data | Clear qualitative properties. Importance familiar from parallel computing and important separately for data and model |
| 9M | D=Dynamic or S=Static Model |
| 10D | R=Regular or I=Irregular Data |
| 10M | R=Regular or I=Irregular Model |
| 11M | Iterative or not? | Clear qualitative property of Model. Highlighted by Iterative MapReduce and always present in classic parallel computing |
| 12D | Data Abstraction | e.g. key-value, pixel, graph, vector, bags of words or items. Clear quantitative property although important data abstractions not agreed upon. All should be supported by Programming model and run time |
| 12M | Model Abstraction | e.g. mesh points, finite element, Convolutional Network. |
| 13D | Data in Metric Space or not? | Important property of data. |
| 13M | Model in Metric Space or not? | Often driven by data but model and data can be different here |
| 14M | O(N2) or O(N) Complexity? | Property of Model algorithm |
|  | | |
| Data Source and Style View of Diamonds (No model involvement except in 9) | | |
| 1D | SQL/NoSQL/NewSQL? | Can add NoSQL sub-categories such as key-value, graph, document, column, triple store |
| 2D | Enterprise data model | e.g. warehouses. Property of data model highlighted in database community / industry benchmarks |
| 3D | Files or Objects? | Clear qualitative property of data model where files important in Science; objects in industry |
| 4D | File or Object System | HDFS/Lustre/GPFS. Note HDFS important in Apache stack but not much used in science |
| 5D | Archived or Batched or Streaming | Streaming is incremental update of datasets with new algorithms to achieve real-time response; Before data gets to compute system, there is often an initial data gathering phase which is characterized by a block size and timing. Block size varies from month (Remote Sensing, Seismic) to day (genomic) to seconds or lower (Real time control, streaming) |
| Streaming Category S1) | S1) Set of independent events where precise time sequencing unimportant. |
| Streaming Category S2) | S2) Time series of connected small events where time ordering important. |
| Streaming Category S3) | S3) Set of independent large events where each event needs parallel processing with time sequencing not critical |
| Streaming Category S4) | S4) Set of connected large events where each event needs parallel processing with time sequencing critical. |
| Streaming Category S5) | S5) Stream of connected small or large events to be integrated in a complex way. |
| 6D | Shared and/or Dedicated and/or  Transient and/or Permanent | Clear qualitative property of data whose importance is not well studied. Other characteristics maybe needed for auxiliary datasets and these could be interdisciplinary, implying nontrivial data movement/replication |
| 7D | Metadata and Provenance | Clear qualitative property but not for kernels as important aspect of data collection process |
| 8D | Internet of Things | Dominant source of commodity data in future. 24 to 50 Billion devices on Internet by 2020 |
| 9 | HPC Simulations generate Data | Important in science research especially at exascale |
| 10D | Geographic Information Systems | Geographical Information Systems provide attractive access to geospatial data |
|  | | |
| Processing (runtime) View of Diamonds | | |
| Big Data and Simulation Processing Kernels | | |
| 1M | Micro-benchmarks | Important subset of small kernels |
| 2M | Local Analytics or Informatics or Simulation | Executes on a single core or perhaps node and overlaps Pleasingly Parallel |
| 3M | Global Analytics or Informatics or simulation | Requiring iterative programming models across multiple nodes of a parallel system |
| 12M | Linear Algebra Kernels | Important property of some analytics |
| Many important subclasses | Conjugate Gradient, Krylov, Arnoldi iterative subspace methods |
| Full Matrix |
| Structured and unstructured sparse matrix methods |
| 13M | Graph Algorithms | Clear important class of algorithms – often hard |
| 14M | Visualization | Clearly important aspect of analysis in simulations and big data analyses |
| 15M | Core Libraries | Functions of general value such as Sorting, Math functions, Hashing |
|  | | |
| Big Data Processing Diamonds | | |
| 4M | Base Data Statistics | Describes simple statistical averages needing simple MapReduce in problem architecture |
| 5M | Recommender Engine | Clear type of big data machine learning of especial importance commercially |
| 6M | Data Search/Query/Index | Clear important class of algorithms – especially in commercial applications. |
| 7M | Data Classification | Clear important class of big data algorithms |
| 8M | Learning | Includes deep learning as category |
| 9M | Optimization Methodology | Includes Machine Learning, Nonlinear Optimization, Least Squares, expectation maximization, Dynamic Programming, Linear/Quadratic Programming, Combinatorial Optimization |
| 10M | Streaming Data Algorithms | Clear important class of algorithms associated with Internet of Things. Can be called DDDAS Dynamic Data-Driven Application Systems |
| 11M | Data Alignment | Clear important class of algorithms as in BLAST |
|  | | |
| Simulation (Exascale) Processing Diamonds | | |
| 16M | Iterative PDE Solvers | Jacobi, Gauss Seidel etc. |
| 17M | Multiscale Method? | Multigrid and other variable resolution approaches |
| 18M | Spectral Methods | Fast Fourier Transform |
| 19M | N-body Methods | Fast multipole, Barnes-Hut |
| 20M | Particles and Fields | Particle in Cell |
| 21M | Evolution of Discrete Systems | Electrical Grids, Chips, Biological Systems, Epidemiology. Needs ODE solvers |
| 22M | Nature of Mesh if used | Structured, Unstructured, Adaptive |