Data Operation, Global Development Operations, Global Drug Development

SAS and R

Sri Vithala, Director
Hyderabad
May 18, 2019

NOVARTIS

Clinical Data Analysis

1. SAS has served as standard with Version 5.0 in public Domain
   - Higher versions introduced more sophistication like PROC MIXED. IML, PHREG, etc.
   - Better handling large data sets
2. New platforms have arrived as Open Source such as Programming Language R and packages
   - lme4, nlme, etc.
   - Plotting abilities
3. Spark & Hadoop for Big Data have R extensions
4. Both SAS and R are trusted by tradition

SAS and/or R for Data Set creation

- R can create CDISC compliant clinical trial datasets
- R is obviously cost effective
- Excellent data visualization - R-Shiny
- R can output SAS7BDAT or CSV file
- No Log file in R
- Upkeep required for updates in R packages
- Cannot add labels to Variables or Data sets

- SAS is validated (SDLC)
  - There are companies offering validated distributions of R.
- SAS has log file
- SAS can add labels to variables/datasets.
- SAS releases have downward compatibility

High Performance Statistical Computing (HPSC)

- Computers used a 32-bit architecture: The maximal amount of memory a process can address (the address space) is 2^{32} bytes, or 4 GB
- More recently 64-bit architecture: Maximal address space is $2^{64} \approx 10^{19} = 10^{7}$TB = 10^{4}PB = 10EB
- Large Data sets require larger amounts of memory using multiple computers
- Multiple standard processors with shared or distributed memory
  - Beowulf clusters (retail)
  - Distributed shared memory systems (proprietary)

High Performance Statistical Computing (HPSC)

- Moore's law does not apply to speed; limited by
  - Heat
  - Power Consumption
- Parallel programming is likely to become essential even for desktop computers
- A simple model says a computation runs N times faster when split over N processors.
- More realistically, a problem has a fraction S of its computation that is inherently sequential and 1 − S that can be parallelized.
- Amdahl's law:

$$\text{Maximum Speedup} = \frac{1}{S+(1-S)/N}$$

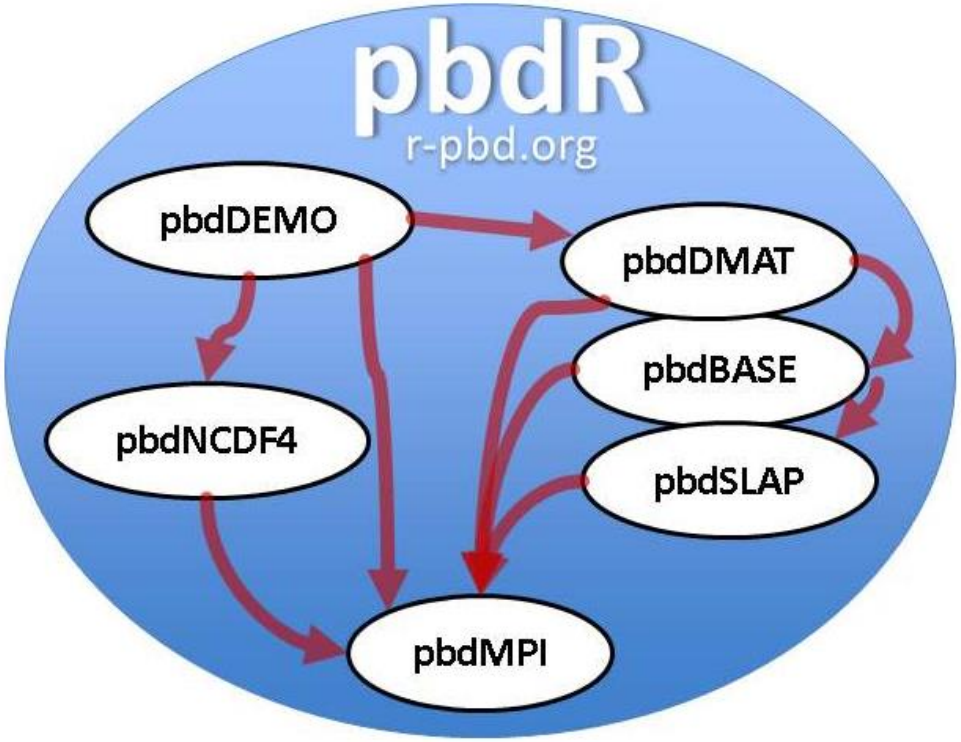
S≈ 0 makes it *embarrassingly parallel*

High Performance Statistical Computing (HPSC)

- Some statistical problems are (or seem to be) parallel:
  - computing column means
  - Bootsrapping & cross-validation, involve independent repeated computations with slightly different inputs
- Others seem inherently sequential:
  - pseudo-random number generation
  - Markov chain Monte Carlo
- Using parallel computing we can speed up nearly linear in the number of processors
  - small deviation from linearity could happen due to communication costs and slight variability in the run times of the individual jobs.

Big Data Programming using R

- **p**rogramming with **b**ig **D**ata in **R** (pbDR) package
- The use of high performance Statistics computation



Differences between R and pbdR

- pbdR mainly focuses on distributed memory systems (processors),
  - data are distributed across several processors
  - analyzed in batch mode
  - high-performance computing (HPC) systems
- communications between processors are based on message-passing interface (MPI)
- CRAN R system focuses mainly on single multi-core machines for data analysis via an interactive mode such as GUI interface.

Big Data Computation with parallel R

High Performance Statistical Computation

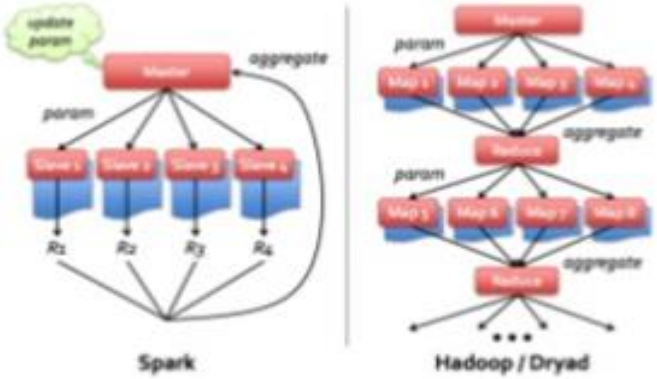
- single program multiple data (SPMD)
  - Ease programming and debugging efforts.
  - Scalability for parallel computing.
  - Distributed system for ultra-large/unlimited datasets.
  - Reduce traffics of communication.
- **Advantages of SPMD over Master/Worker**
  - is easy to modify from serial.
  - is traceable for debugging.
  - fully utilizes resources.
  - can do interaction, as well.
  - can deal with ultra-large/unlimited data.
  - can parallelize by jobs.

Big Data Management

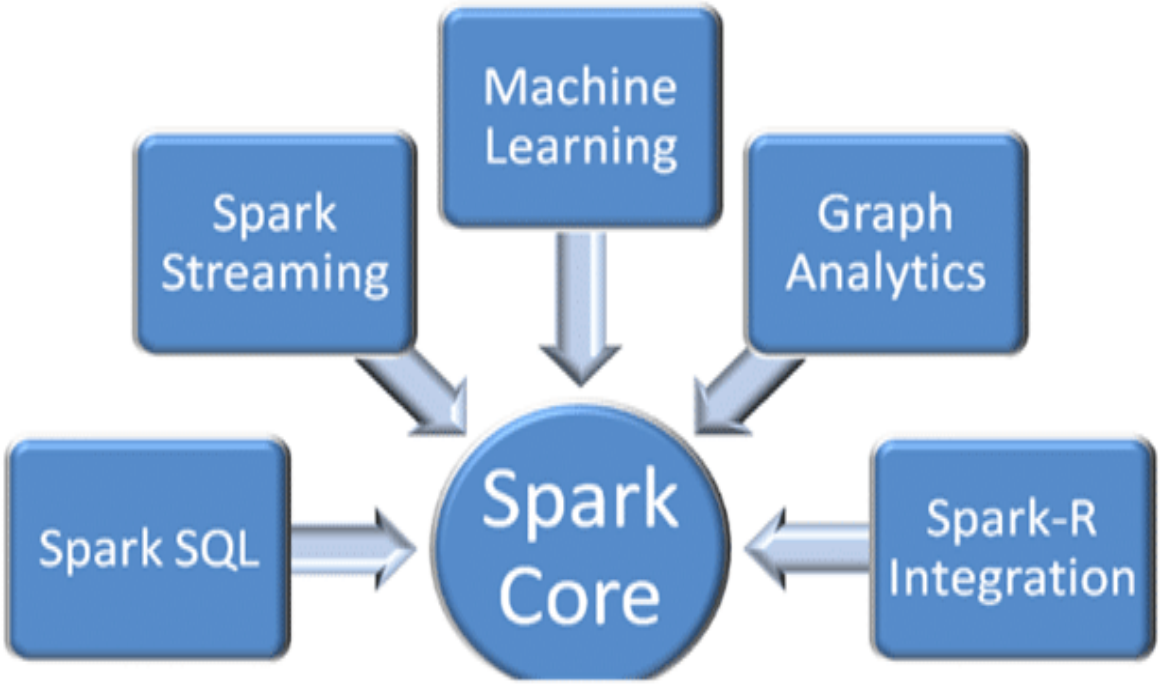
1. Couple of models
   - Hadoop MapReduce
   - Spark

Spark vs Hadoop MapReduce

- In-memory data flow model optimized for multi-stage jobs
- Novel approach to fault tolerance
- Similar programming style to Scalding/Cascading



Big Data Extensions for R



NOVARTIS

Real World Evidence with Big data

- “Big data” can be considered the material basis for the realization of RWE studies. These may be a complementary partner of RCTs and thereby a valuable tool in clinical research in oncology.
- Modern IT concepts and technologies enable the digital and structured capture of complex oncological information in addition to routine medical data.
- Thereby, it is possible to analyze data longitudinally and data that has been collected with different methods.
- The individual centers in a RWE study network may conduct national or international benchmarking, depending on the composition of the network, in addition to analyzing their internal clinical context.
- This may not only yield clues about outcomes of current treatment pathways but also about alternative approaches or about new research-related hypotheses.
- RWE studies are therefore a meaningful complement to RCTs which typically analyze pre-selected patients and to clinical registries which usually operate on a reduced data scope.
- The ambition to conduct high-quality RWE studies in oncology poses significant challenges for all stakeholders with regards to IT, personnel, organizational, financial, and data privacy aspects.
- These challenges can only be overcome jointly in order to achieve the legitimate aim of a relevant improvement of quality, effectiveness and safety in oncological care.

SAS Salient features

- Commercial software with licensing costs
- Easy to learn (like SQL)
- Graphing capabilities are OK.
  - Not a match to R
- Extensive libraries and wide-ranging documentation
  - Also in R
- Market leader for corporates due to software validation & History
- SAS Proc IML can be used to submit R code
  - Reads data in Matrix structure

SAS Analytics for Big Data

- The SAS LASR Analytic Server
  - secure, multi-user environment for concurrent access to data that is loaded into memory.
  - Can access a distributed computing environment for parallel processing
  - The server handles both big data and smaller sets of data
  - Designed with a high-performance, multi-threaded, analytic code
  - Performance at high speeds due to the combination of hardware and software
  - Rapid access to tables in memory.
    - Tables loaded into memory for analytic processing
  - HDFS acts as a co-locator data provider
  - Machine running SAS LASR Analytic Server act as NFS clients to mount the MapR cluster.
    - Data Loader can be used to Load data to LASR
    - No client software is required
  - Several directives in Data Loader 2.4 support Apache Spark

Data Analytics – ML, AI

- Clinical medicine is handling enormous amounts of data from RWE as well as Data pools like Data Lakes
  - “-omic” data like genomic, etc. are very large
  - Machine learning & Artificial Intelligence will help Clinical trials to be better focused
  - As patients’ conditions and medical technologies become more complex, Deep Learning Algorithms will drive clinical medicine
    - Detection of disease by reviewing images
  - Create winners and losers in medicine
  - Data analysis using underlying algorithms will transforms clinical medicine
- Analysis of large Data Lakes can provide
  - insights into disease characteristics
  - can form hypothesis in basic Clinical Research
  - Pooled data of a drug can drive future approaches in trial designs

Statistical Analysis with R & SAS

Both SAS and R have powerful Statistical Analysis packages

- SAS Procs are developed to be one stop shop with options and statements
- R Packages are also developed with many features to give the users to set various statistical criteria
- SAS develops tools and procs at each release
- R packages can be released at any time by any user/developer/enthusiast
- NOTE: Python can reproduce much of the Statistical Analysis as in R or SAS but too cumbersome for extremely complex analysis which are easily done in R or SAS

Statistical Analysis with R & SAS

- R Packages are developed faster than SAS

R	SAS
GLMM Package	Proc GLMMIX
Ada booster Package	No procedure (Macros available)

- Working in SAS and R can produce different results
  - For example in Mixed models specifics around degrees of freedom can be different
  - Nesting could require Laplace approximation or more general Gauss-Hermite Quadrature in both Proc Mixed (SAS) and lme4 (R) for maximum likelihood
  - More code and effort to create R plots that are Similar to SAS, however, overall better plots in R

Analysis in SAS and Validating in R

- Most regulatory submission are happening in still in SAS
- The submitted Data Sets can be reviewed by FDA and others
- Regulatory agencies see value in Direct Data Review
- Advisory committees are asked to evaluate the risk/benefit of drugs based on independent Data Review
- Best to have data in SAS and R and done some basic validation to ensure results are consistent and rubust

Analysis in SAS and Validating in R

- FDA has created Office of Drug Evaluation Sciences
- Data Review will be in Cloud based environment
- It will also evaluate elements that are having an increasing role in the drug review process
- Biomarkers, bioinformatics, patient-reported outcomes, etc. are some of the newer elements
- The office will be staffed to perform their own analysis
- Possible tools like R and Rshiny, or perhaps continue use of SAS/JMP
- Of course, it is important that they receive CDISC Standard data
- Expert teams would be in place to do this review
- The group will begin this year with a structured review of safety data

Thank you

NOVARTIS