

天文学の大規模データへの応用

高深度宇宙観測データ処理DB

高田唯史、古澤久徳、古澤順子、大坪貴文
(国立天文台天文データセンター)

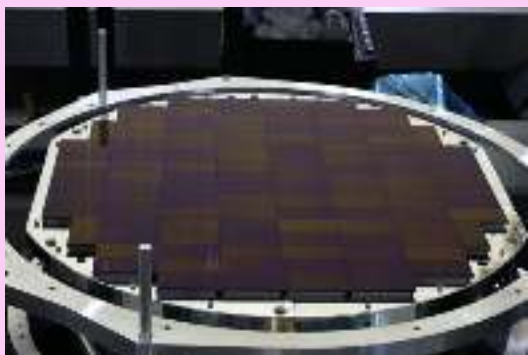
高速データベースと天文学の大規模データ

- 口径8メートルの光学望遠鏡（すばる）に装着された超広視野カメラ（Hyper Suprime-Cam:HSC）による深宇宙探査観測による宇宙の進化の研究
 - ◆1000平方度を超える広い視野の観測（**視野数は1000以上**）
 - ◆各露出により**50～100万個程度の天体**（+ α ）に関する明るさ、位置、形の測定値
 - ◆測定値に関連する**パラメータは数百～数千**
 - ◆**露出数は10～数百程度**（同じ天体に対する時系列情報）
 - ◆波長方向には**5つ（または10以上）のフィルター**
- 高速検索する必要のあるデータベーステーブルの規模は
レコード数：（観測する視野の数）×（露出あたりの天体数）×（露出数）×（フィルター数）
カラム数：測定パラメータ数
 - ✓8億以上の天体に対して**0（ 10^3 ）個のカラム×5種類以上の波長情報**
 - ✓**300～400億天体（延べ数）で1000個程度のカラム×5種類の波長情報**のようなデータベーステーブルに対する高速検索＋高速分析が必要になる。

約1.7度

30分角

1領域で約13億
ピクセルの画像



超広視野カメラ HSC



すばる
望遠鏡

約15分角

約2分角 (0.033度)



-3.07e-13 -9.46e-14 1.20e-13 3.32e-13 5.46e-13 7.58e-13 9.70e-13 1.18e-12 1.40e-12

データ量(HSC戦略枠5年分(1世代)の試算)

各サーベイの生み出すデータプロダクトのカタログデータ量(20160629 版)

サーベイ	CCD 画像数	CCD カタログ 総天体数	CCDDB 容 量(GB)	モザイク 画像数	モザイクカタ ログ総天体数	モザイク DB 容量(GB)
Wide	2445410	35702993291	71400	150000	13000000000	62100
Deep	355158	5185302959	10400	44800	500000000	2000
Ultra-Deep	205148	2995161184	6000	1200	100000000	600
合計	3005716	43883457434	87800	196000	13600000000	64700

HSC 戦略枠サーベイと既存サーベイのデータ量比較

	画像ファイル数	画像ファイル容量 (TB)	カタログ天体数	カタログ DB 容量 (GB)
HSC 戦略枠	220000000	900.0	13600000000	152500
(フレーム毎)			439000000000	1219
SDSS-DR7	65000000	15.7	3570000000	2600
2MASS	4121439	8.5	4710000000	160

**総天体数 ~13.6億(~2000カラム)
延べ数(439億)(数百カラム)**

生データ 2GB/shot

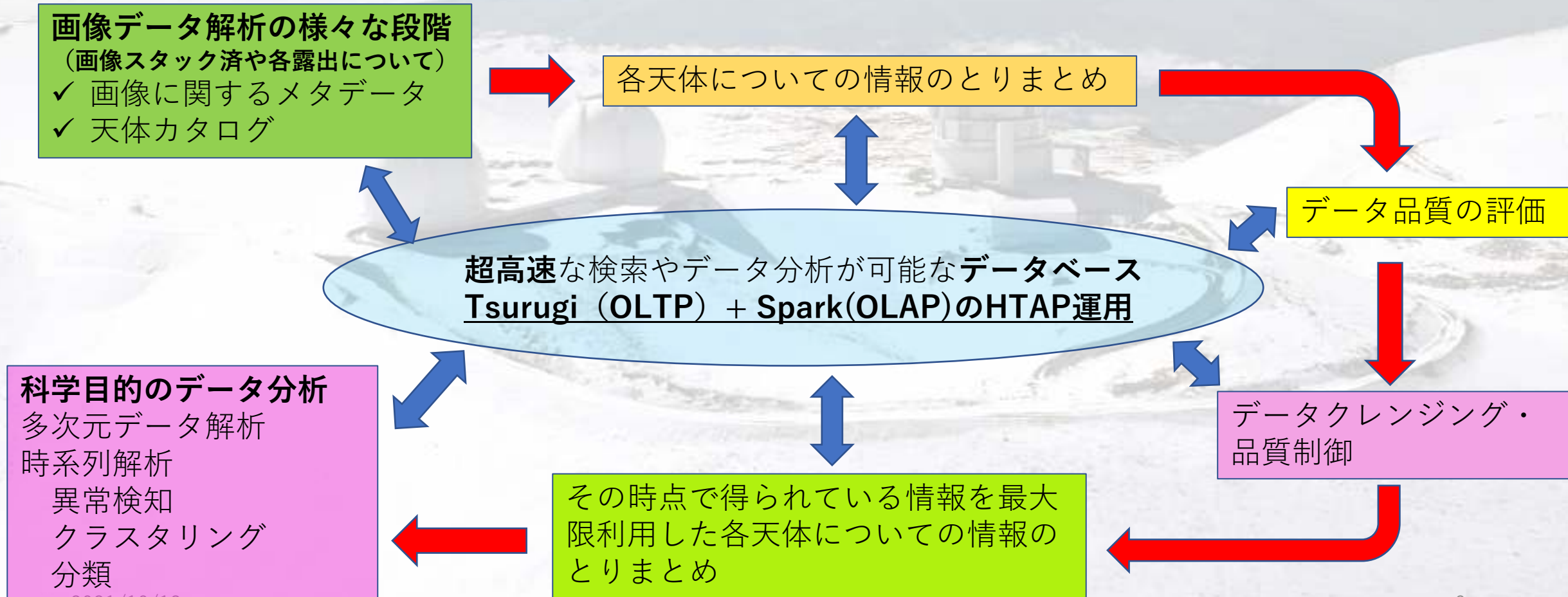
**画像容量 900TB+ α
データベース 152TB+ α ?**

Tsurugiプロジェクトにおける天文データの応用

-- 2つの柱 --

- (最新かつ大量なデータを用いた) 複雑なクエリワークロードを利用したスキーマ推薦と大規模データ分析
 - ◆現在のPostgreSQLクラスターでは遅い(遅すぎる)クエリーの高速化
 - ✓ 12時間以上かかるようなクエリーを高速化して、研究の高効率化
 - ✓ クエリーログ解析によるクエリー高速化情報の構築(スキーマ推薦)
 - ✓ 画像解析の各段階で生産される多様な情報の連携による高度なスクリーニング(前処理)
 - ✓ 今までできなかった検索結果を用いたトライ&エラー・チャレンジ→データ品質評価・管理・制御
 - ✓ 発展型としてのマルチバンドカタログを利用した距離推定の高速化と銀河進化の研究
- 大量のデータを用いた探索的データ分析
 - ◆巨大テーブルを用いての時間軸方向の探索的データ分析
 - ✓ 時系列データ作成による変光天体や移動天体の検出、それによる科学研究
 - ✓ 数十から数百回の観測・測定→レコード数が多すぎて既存DBエンジンでは苦しい
 - ✓ 変光天体(明るさの変化する天体)の(自動)分類
 - ✓ 多数(数百万超)の巨大ブラックホール天体による観測的宇宙進化の研究
 - ✓ 重力波天体・超新星等の突発天体検出と即時アラート発信システムにつなげる
 - ✓ 移動天体の検出→既存天体の軌道情報改善、新天体の発見(危険小惑星) + α

高速データベース導入による データ戦略スキーム



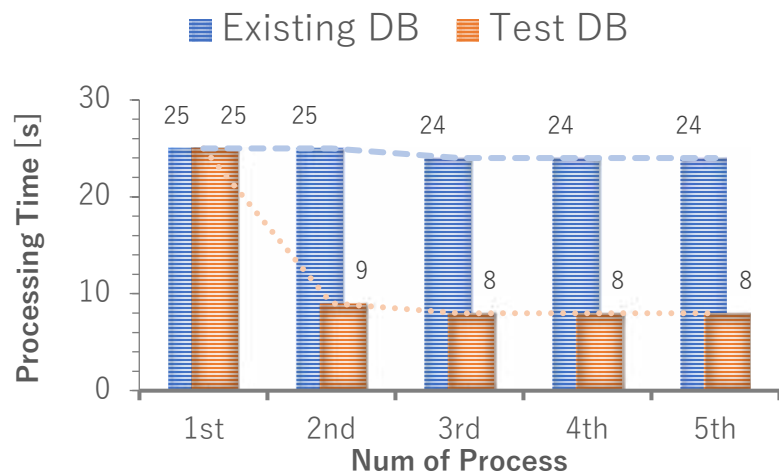
OLAP：検索時間の比較

散処理)

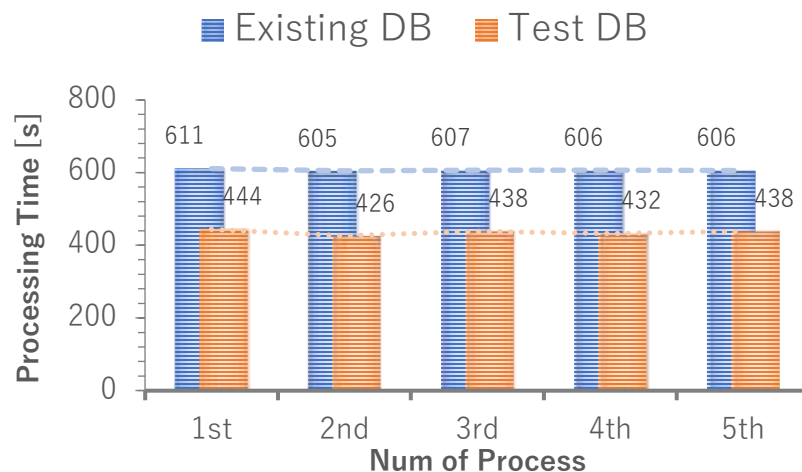
PostgreSQLクラスター（旧来处理） vs Spark環境（分



QUERY 1



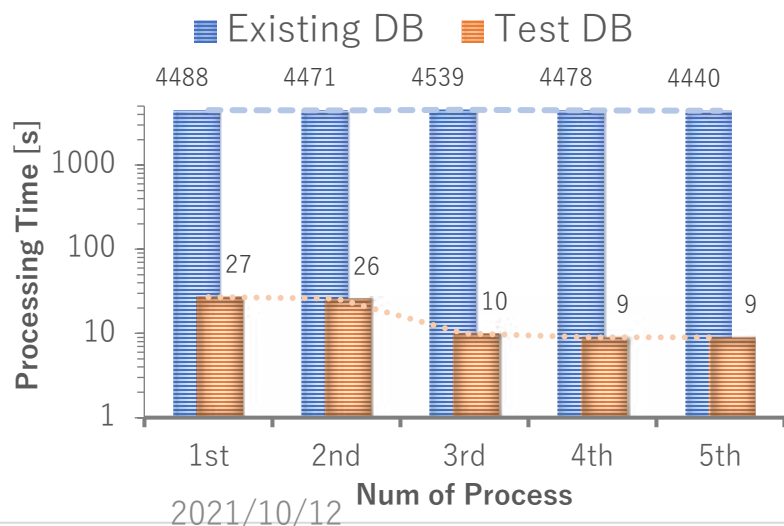
QUERY 2



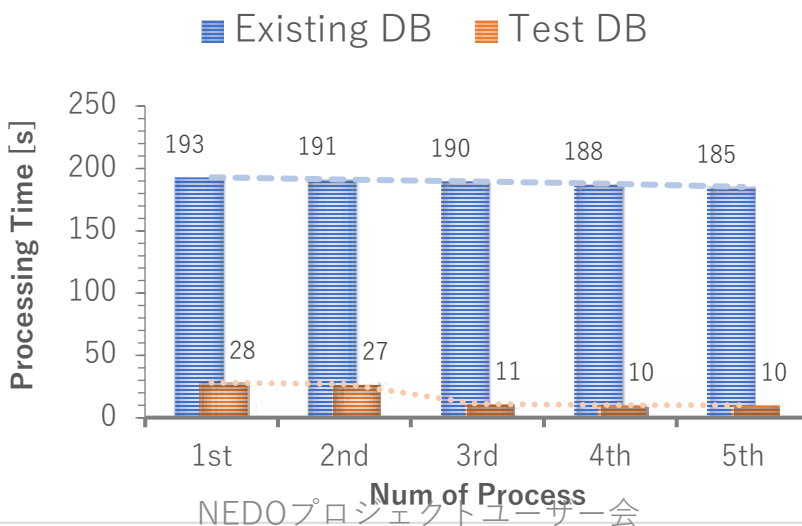
- 計測した全てのqueryにおいてSpark環境(分散処理)の方が処理が高速(query3は2桁の違い)
- 初回の処理時間のみに注目した場合、ほぼ差がないqueryもあり(query1)

QUERY 3

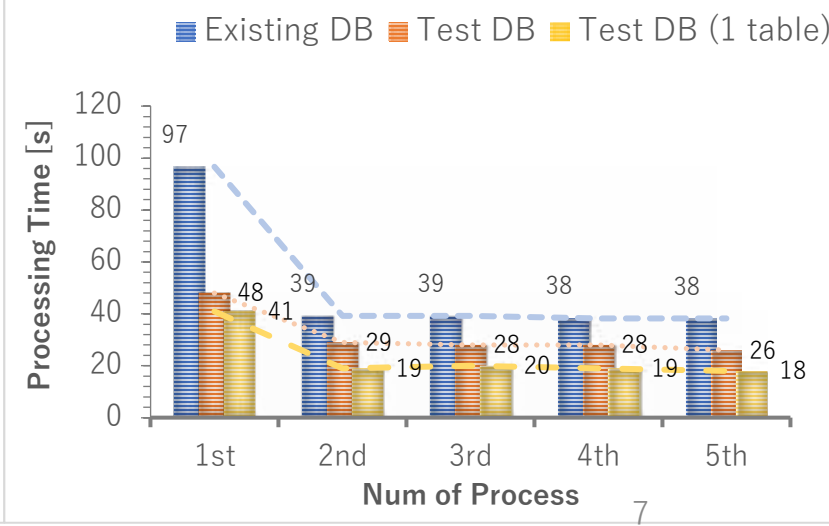
2桁速い！



QUERY 4



QUERY 17



Innovative astronomical applications with a new-generation relational database

Junko Furusawa^{*a}, Tadafumi Takata^{a,b}, Hisanori Furusawa^{a,b}, Takafumi Ootsubo^a, Shohei Aoyama^c,
Yoshihiko Yamada^{a,d}, Yuki Okura^a, Makoto Onizuka^e,

^aNational Astronomical Observatory of JAPAN, National Institutes of Natural Science (NINS), 2-21-1 Osawa, Mitaka, Tokyo, Japan 181-8588; ^bThe Graduate University for Advanced Studies, SOKENDAI, Shonan Village, Hayama, Kanagawa, Japan 240-0193; ^cInstitute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, Japan 277-8583; ^dJapan Space Forum, Shin-Ochanomizu Urban Trinity Bldg. 3F 3-2-1, Kandasurugadai, Chiyoda-ku, Tokyo, Japan 101-0062; ^eGraduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka, Japan 565-0871

ABSTRACT

Database technology has been developing to exploit the next-generation hardware in the era of big data processing. At the same time, astronomical data size has been steadily increasing, and astronomical source catalogs obtained from large-scale surveys with a wide-field camera, such as Subaru/Hyper Suprime-Cam (HSC), are a good test bench for evaluating the new database technology with a large data set. Such archive systems often employ a highly versatile relational database management system (RDBMS), but reducing the time required for data transaction and complex analysis has come to an important challenge. To tackle this difficulty, we aim to develop astronomical applications with a new catalog database using a next-generation RDBMS technology, where the query engine is designed to efficiently use computing infrastructures for processing big data. Demonstrations with science applications are essential to evaluate the new database. We verify query performance with the current HSC source catalog. For application to huge astronomical catalog databases, we are pursuing and verifying the capabilities of new database technologies. It will, in turn, enable fast ad hoc search and efficient detection of a wide range of variable events with the technology. Our pilot tests using typical astronomical queries on a cluster system shows significant improvements in response times with the aid of distributed query engines. We report performance of the test database for typical astronomical queries, and discuss optimizing the schema based on query workloads.

Keywords: Relational Database, Source Catalog, Schema Optimization



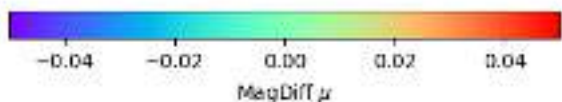
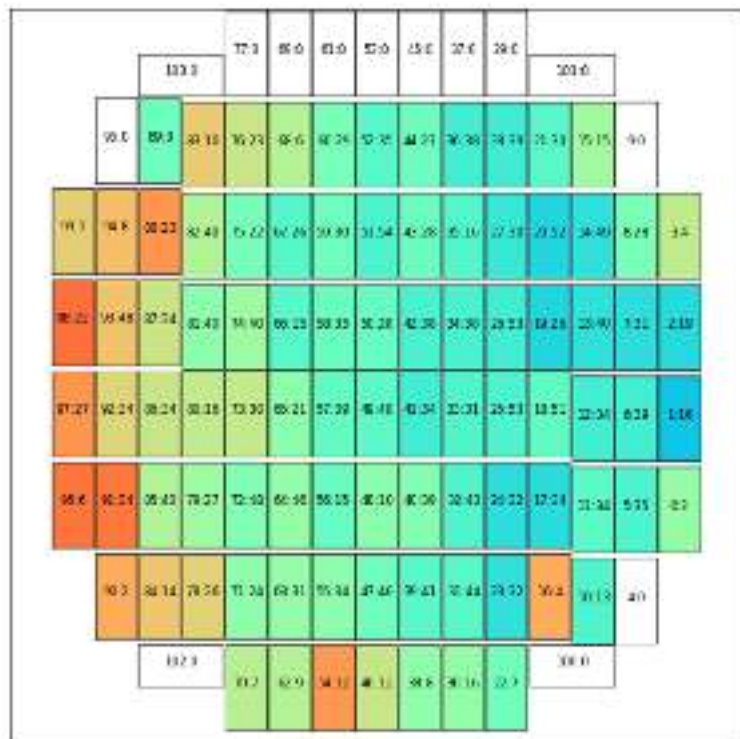
Sparkを用いたクエリー高速化については
SPIE (The International Society for
Optical Engineering) のカンファレンス
“Astronomical Telescopes +
Instrumentation 2020”

(<https://spie.org/conferences-and-exhibitions/astronomical-telescopes-and-instrumentation?SSO=1>) においてJ.
Furusawa et al.(2020)論文として発表

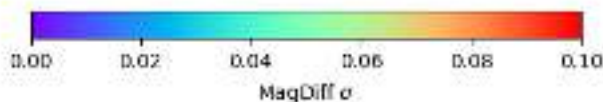
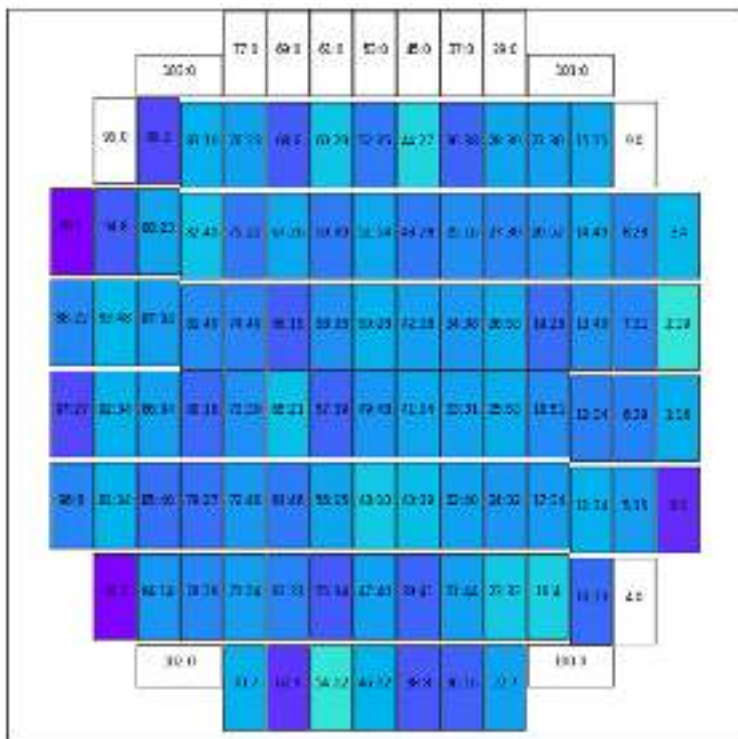
品質評価の一例：

各CCDにまで遡ったデータ品質評価

Tract:9813 Visit:1176 Filter:HSC-Z Magtype:ap40



測光ゼロ点のズレ



明るさの差の分布の標準偏差

測光（天体の明るさを測定すること）の精度の評価

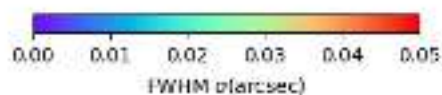
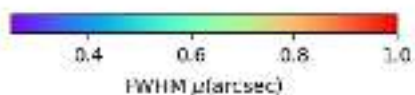
明るさが変動していないと思われる天体をDBからピックアップ

各天体について全てのある程度性とよく測定できたと思われる測定値の平均（メジアンでも良い）とある露出時の明るさの差を求める

各CCDについて情報をまとめ、平均値と標準偏差をマップする

各CCDにまで遡ったデータ品質評価

測定値



各CCDについての星像サイズの評価

各露出について、S/Nの良いデータに基づいた情報を利用して、星と思われる天体を抽出

抽出された各天体について星像サイズの推定

各CCDについてまとめた上で、平均値、分布の標準偏差、それらの空間分布を確認

観点：

あまりに星像が変なデータは分析から除外
測定値の星像（要は地球大気の状態）依存度
空間（2次元）分布（特にモデル）の正常性確認

星像モデル

